TESIS

ANALISA DAN KOMPARASI IMPLEMENTASI DATA MINING UNTUK PREDIKSI KELULUSAN TEPAT WAKTU MAHASISWA

Disusun oleh:

Nama : **Azis Wahyudi**
NIM : **21.52.1050**
Konsentrasi : **Business Intelligence**

PROGRAM STUDI S2 TEKNIK INFORMATIKA
PROGRAM PASCASARJANA UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2024

TESIS

ANALISA DAN KOMPARASI IMPLEMENTASI DATA MINING UNTUK PREDIKSI KELULUSAN TEPAT WAKTU MAHASISWA

ANALYSIS AND COMPARISON OF DATA MINING IMPLEMENTATION FOR PREDICTING STUDENTS' TIMELY GRADUATION

Diajukan untuk memenuhi salah satu syarat memperoleh derajat Magister

Disusun oleh:

Nama : **Azis Wahyudi**
NIM : **21.52.1050**
Konsentrasi : **Business Intelligence**

PROGRAM STUDI S2 TEKNIK INFORMATIKA
PROGRAM PASCASARJANA UNIVERSITAS AMIKOM YOGYAKARTA
YOGYAKARTA
2024

HALAMAN PENGESAHAN

ANALISA DAN KOMPARASI IMPLEMENTASI DATA MINING UNTUK PREDIKSI KELULUSAN TEPAT WAKTU MAHASISWA

ANALYSIS AND COMPARISON OF DATA MINING IMPLEMENTATION FOR PREDICTING STUDENTS' TIMELY GRADUATION

Dipersiapkan dan Disusun oleh

Azis Wahyudi

21.52.1050

Telah Diujikan dan Dipertahankan dalam Sidang Ujian Tesis
Program Studi S2 Teknik Informatika
Program Pascasarjana Universitas AMIKOM Yogyakarta
pada hari Jumat, 02 Februari 2024

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister Komputer

Yogyakarta, 02 Februari 2024
Rektor

Prof. Dr. M. Suyanto, M.M.
NIK. 190302001

HALAMAN PERSETUJUAN

ANALISA DAN KOMPARASI IMPLEMENTASI DATA MINING UNTUK PREDIKSI KELULUSAN TEPAT WAKTU MAHASISWA

ANALYSIS AND COMPARISON OF DATA MINING IMPLEMENTATION FOR PREDICTING STUDENTS' TIMELY GRADUATION

Dipersiapkan dan Disusun oleh

Azis Wahyudi

21.52.1050

Telah Diujikan dan Dipertahankan dalam Sidang Ujian Tesis
Program Studi S2 Teknik Informatika
Program Pascasarjana Universitas AMIKOM Yogyakarta
pada hari Jumat, 02 Februari 2024

Pembimbing Utama

Prof. Dr. Kusrini, M.Kom.
NIK. 190302106

Anggota Tim Penguji

Tonny Hidayat, M.Kom., Ph.D.
NIK. 190302106

Pembimbing Pendamping

Ferry Wahyu Wibowo, S.Si., M.Cs.
NIK. 190302235

Dhani Ariatmanto, M.Kom., Ph.D.
NIK. 190302197

Prof. Dr. Kusrini, M.Kom.
NIK. 190302106

Tesis ini telah diterima sebagai salah satu persyaratan
untuk memperoleh gelar Magister Komputer

Yogyakarta, 02 Februari 2024
Direktur Program Pascasarjana

Dr. Kusrini, M.Kom.
NIK. 190302106

HALAMAN PERNYATAAN KEASLIAN TESIS

Yang bertandatangan di bawah ini,

Nama mahasiswa : Azis Wahyudi
NIM : 21.52.1050
Konsentrasi : Business Intelligence

Menyatakan bahwa Tesis dengan judul berikut:
Analisa dan Komparasi Implementasi Data Mining untuk Prediksi Kelulusan Tepat Waktu Mahasiswa

Dosen Pembimbing Utama : Prof. Dr. Kusrini, M.Kom
Dosen Pembimbing Pendamping : Ferry Wahyu Wibowo, S.Si., M.Cs.

1. Karya tulis ini adalah benar-benar ASLI dan BELUM PERNAH diajukan untuk mendapatkan gelar akademik, baik di Universitas AMIKOM Yogyakarta maupun di Perguruan Tinggi lainnya
2. Karya tulis ini merupakan gagasan, rumusan dan penelitian SAYA sendiri, tanpa bantuan pihak lain kecuali arahan dari Tim Dosen Pembimbing
3. Dalam karya tulis ini tidak terdapat karya atau pendapat orang lain, kecuali secara tertulis dengan jelas dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan disebutkan dalam Daftar Pustaka pada karya tulis ini
4. Perangkat lunak yang digunakan dalam penelitian ini sepenuhnya menjadi tanggung jawab SAYA, bukan tanggung jawab Universitas AMIKOM Yogyakarta
5. Pernyataan ini SAYA buat dengan sesungguhnya, apabila di kemudian hari terdapat penyimpangan dan ketidakbenaran dalam pernyataan ini, maka SAYA bersedia menerima SANKSI AKADEMIK dengan pencabutan gelar yang sudah diperoleh, serta sanksi lainnya sesuai dengan norma yang berlaku di Perguruan Tinggi

Yogyakarta, 02 Februari 2024
Yang Menyatakan,

Azis Wahyudi

HALAMAN PERSEMBAHAN

Tesis ini penulis persembahkan kepada :

1. Sebagai niat ibadah kepada Allah Subhana Wata'ala, Insya Allah diselesaikan dengan hati tulus ikhlas dipersembahkan kepada para penghaus dan pecinta ilmu semoga menjadi ilmu yang bermanfaat.
2. Ibu Tri Umi Bidayah yang selalu memberikan dorongan semangat, doa, senantiasa menjadi alarm alami pengingat untuk menyelesaikan tesis tepat waktu serta menjadi salah satu alasan utama tetap fokus dalam proses penelitian tesis.
3. Seluruh dosen dan karyawan Magister Teknik Informatika Universitas AMIKOM Yogyakarta.
4. Seluruh civitas akademik angkatan 21 B Magister Teknik Informatika Universitas AMIKOM Yogyakarta.
5. Pembaca yang budiman.

KATA PENGANTAR

Segala puji penulis penjatkan kepada Allah Subhana Wata'ala atas segala limpahan rahmat dan karunia-Nya, sehingga penulis dapat menyelesaikan Tesis ini dengan judul — Analisa dan Komparasi Implementasi Data Mining untuk Prediksi Kelulusan Tepat Waktu Mahasiswa ini. Penyusunan laporan ini tidak lepas atas bimbingan dan sumbangsih dari berbagai pihak. Oleh karena itu dalam kesempatan ini penulis mengucapkan terimakasih kepada :

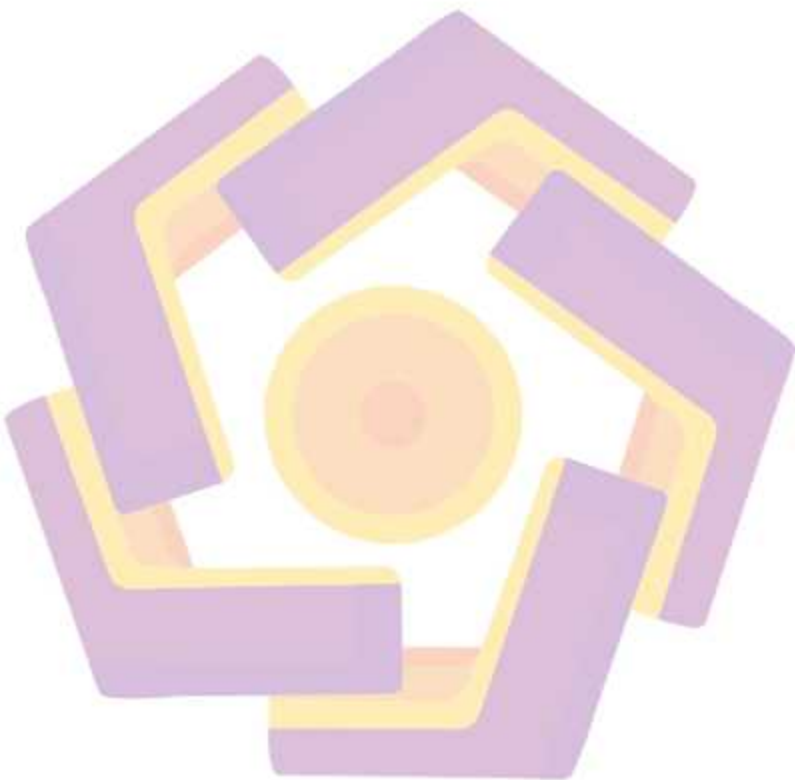
1. Bapak Prof. Dr. M. Suyanto, M.M. selaku Rektor Universitas Amikom Yogyakarta yang berkenan memberika kesempatan untuk menimba ilmu di Universitas AMIKOM Yogyakarta ini.
2. Ibu Prof. Dr. Kusrini, M.Kom. selaku dosen pembimbing utama dan bapak Ferry Wahyu Wibowo, S.Si., M.Cs. selaku dosen pembimbing pendamping yang telah banyak memberikan ilmu, waktu, dan segenap perhatiannya selama membimbing saya dalam menyelesaikan tesis ini.
3. Orang tua, Istri, beserta kakak dan adik yang selalu mendukung dan memberikan doa dan motivasinya.
4. Rekan-rekan mahasiswa Universitas AMIKOM Yogyakarta yang telah berjuang bersama menyelesaikan studi S2.
5. Segenap Dosen Universitas AMIKOM Yogyakarta yang telah banyak memberikan ilmu selama menimba ilmu dikampus ini
6. Terakhir kepada semua pihak yang telah membantu, yang tadak dapat diuraikan satu persatu.

Penulis menyadari bahwa masih ada langit diatas langit, dan begitu juga dengan Tesis ini yang masih sangat perlu disempurnakan dan dikembangkan lagi. Oleh karena itu, penulis membuka diri untuk saran dan kritik yang membangun atas nama ilmu pengetahuan. Penulis juga berhadap bahwa tesis ini dapat dijadikan rujukan ataupun sumbangsi terhadap ilmu pengetahuan untuk dikemudian hari.

Yogyakarta, 02 Februari 2024

Penulis

DAFTAR ISI



DAFTAR TABEL

DAFTAR GAMBAR

INTISARI

Perguruan tinggi memiliki peran penting dalam menyediakan pengetahuan yang diperlukan mahasiswa sebelum memasuki dunia kerja. Keberhasilan sebuah universitas, baik yang negeri maupun swasta, sering kali diukur dari jumlah mahasiswa yang lulus tepat waktu. Penelitian ini bertujuan untuk mengevaluasi implementasi algoritma C4.5, Klasifikasi Naive Bayes, dan K-Nearest Neighbors dalam memodelkan prediksi kelulusan tepat waktu, dengan fokus pada pengukuran akurasi prediksi menggunakan teknik n-Folds Cross Validation. Data akademis yang digunakan dalam penelitian ini mencakup sampel mahasiswa dari Poltekkes Permata Indonesia pada Tahun Akademik 2019/2020. Pendekatan eksperimental diterapkan, di mana metode Klasifikasi C4.5 dibandingkan dengan Naive Bayes dan K-Nearest Neighbors.

Proses pelatihan dan pengujian sistem dilakukan menggunakan metode 5-fold Cross Validation, dengan fokus pada pengukuran akurasi, presisi, dan recall sebagai hasil evaluasi. Hasil pengujian kinerja ketiga metode ini menggunakan cross-validation, matriks kebingungan, dan kurva ROC menunjukkan bahwa C4.5 memiliki akurasi tertinggi, mencapai 85,24%, dengan presisi sebesar 96,03%, recall sebesar 87,32%, dan F1 Score sebesar 91,49%.

Pada urutan kedua, K-Nearest Neighbors (KNN) mencapai tingkat akurasi sebesar 84,26%, recall sebesar 86,17%, presisi sebesar 96,43%, F1 Score sebesar 91,01%, dan nilai AUC sebesar 84,6%. Terakhir, Naive Bayes (NB) mencatat tingkat akurasi sebesar 81,31%, recall sebesar 90,41%, presisi sebesar 86,51%, F1 Score sebesar 88,41%, dan nilai AUC sebesar 81,3%. Temuan ini memberikan wawasan tentang efektivitas metode data mining dalam meramalkan kelulusan mahasiswa, dengan C4.5 menunjukkan kinerja optimal dalam konteks ini.

Kata kunci: Data Mining, algoritma C4.5, Naive Bayes, K-Nearest Neighbors.

ABSTRACT

The higher education sector plays a crucial role in providing the necessary knowledge for students before entering the workforce. The success of a university, whether public or private, is often measured by the number of students who graduate on time. This research aims to evaluate the implementation of the C4.5 algorithm, Naive Bayes classification, and K-Nearest Neighbors in modeling timely graduation predictions, focusing on measuring prediction accuracy using n-Folds Cross Validation technique. The academic data used in this research includes samples of students from Poltekkes Permata Indonesia in the Academic Year 2019/2020. An experimental approach is applied, where the C4.5 Classification method is compared with Naive Bayes and K-Nearest Neighbors.

The training and testing process of the system is conducted using the 5-fold Cross Validation method, with a focus on measuring accuracy, precision, and recall as evaluation results. The performance testing results of these three methods using cross-validation, confusion matrices, and ROC curves show that C4.5 has the highest accuracy, reaching 85.24%, with a precision of 96.03%, recall of 87.32%, and F1 Score of 91.49%.

In the second place, K-Nearest Neighbors (KNN) achieved an accuracy rate of 84.26%, recall of 86.17%, precision of 96.43%, F1 Score of 91.01%, and AUC value of 84.6%. Lastly, Naive Bayes (NB) recorded an accuracy rate of 81.31%, recall of 90.41%, precision of 86.51%, F1 Score of 88.41%, and AUC value of 81.3%. These findings provide insights into the effectiveness of data mining methods in predicting student graduation, with C4.5 demonstrating optimal performance in this context.

Keywords: Data Mining, C4.5 algorithm, Naive Bayes, K-Nearest Neighbors.

BAB I

PENDAHULUAN

1.1. Latar Belakang Masalah

Kelulusan merupakan salah satu komponen bagi dalam penilaian akreditasi. Mahasiswa lulus tepat waktu maka akan membantu penilaian akreditasi terhadap program studi hingga perguruan tinggi. (BANPT, 2007) Kelulusan tepat waktu sendiri merupakan salah satu indikator keberhasilan hasil kinerja akademik mahasiswa. Ketentuan masa studi sendiri sudah diatur dalam ketetapan Peraturan Menteri dan Pendidikan Kebudayaan Indonesia yang menjelaskan bahwa kompetensi lulusan bagi mahasiswa diploma tiga dapat menyelesaikan beban wajib minimal 108 SKS dengan masa studi 3 (tiga) tahun dan paling lama 5 (lima) tahun akademik untuk program diploma tiga. (Permenristekdikti No. 44 Tahun 2015)

Namun waktu kelulusan mahasiswa tidak selalu dapat dideteksi secara dini khususnya di poltekkes permata Indonesia yogyakarta, sehingga bisa mengakibatkan keterlambatan lulusan. Untuk mengatasi hal tersebut perlu ada teknik untuk bisa melakukan prediksi terhadap kelulusan. Adapun teknik yang sering digunakan adalah dengan menggunakan data mining. Dan metode yang sering digunakan untuk prediksi kelulusan mahasiswa adalah metode klasifikasi. (Rohmawan, 2018) Menurut (Suntoro, 2019) data mining adalah proses untuk mendapatkan informasi yang berguna dari basis data yang besar dan perlu diekstraksi agar menjadi informasi baru dan dapat membantu dalam pengambilan keputusan. Data mining adalah proses menganalisa data dari yang berbeda dan

menyimpulkannya menjadi informasi atau pengetahuan atau pola yang penting untuk meningkatkan keuntungan, memperkecil biaya pengeluaran, atau bahkan keduanya (Witten, 2016).

Data mining sudah ada sejak lama dan teori-teorinya pun sudah banyak dibahas dalam literatur. Teori-teori tersebut antara lain: Naive-Bayes dan Nearest Nighbour, Pohon Keputusan, Aturan Asosiasi, k-Means Clustering, dan Text Mining (Bramer, 2007). Metode klasifikasi merupakan pendekatan untuk menjalankan fungsi klasifikasi dalam data mining yaitu menggolongkan data. Teknik klasifikasi ini dapat pula digunakan untuk melakukan prediksi atas informasi yang belum diketahui sebelumnya. Beberapa algoritma yang dapat digunakan antara lain adalah algoritma Decision Tree C.45, Artificial Neural Networks (ANN), K-Nearest Neighbour (KNN), Algoritma Naive Bayes, Algoritma Genetik, Rough Set, Metode Berbasis Aturan, Memory Based Reasoning, dan Support Vector Machine (Widodo dkk., 2013).

Penelitian yang menggunakan data mining pada data set akademik dan kemahasiswaan telah banyak dilakukan, antara lain adalah penelitian yang dilakukan (Nabila. dkk, 2021) yang berjudul "Model Prediksi Kelulusan Tepat Waktu Dengan Metode Fuzzy CMeans Dan K-Nearest Neighbors Menggunakan Data Registrasi Mahasiswa"

Penelitian ini merupakan penggabungan antara algoritma Fuzzy C-Means dan KNearest Neighbors, menggunakan bahasa pemrograman python dengan tools Jupyter Notebook, dataset yang digunakan data registrasi mahasiswa UINSA,

kemudian pengujian skor akan digunakan confusion matrix dan k-fold cross validation. Hasil dari algoritma FCM-KNN didapatkan bahwa model prediksi dengan pengujian 10-fold cross validation dengan skenario k=1 mempunyai rata rata akurasi sebesar 71% .

Penelitian yang dilakukan (Mulia dan Muanas, 2021) "Model Prediksi Kelulusan Mahasiswa Menggunakan Decision Tree C4.5 dan Software Weka" Dalam penelitian ini dibangun sebuah model untuk memprediksi status kelulusan mahasiswa Institut Bisnis dan Informatika Kesatuan menggunakan algoritma pohon keputusan C4.5. Model prediksi dibangun dengan menggunakan IPK mahasiswa semester 1 sampai semester 4, untuk mahasiswa tahun masuk 2013 sampai 2016. Model prediksi yang didapat adalah pohon keputusan dengan 26 aturan, dengan atribut IPS_4 menjadi atribut yang menentukan label kelulusan dari siswa. Model prediksi ini menghasilkan akurasi sebesar 73%, hasil yang kurang baik. Hasil ini kemungkinan disebabkan oleh proporsi data yang digunakan tidak seimbang.

Penelitian serupa dilakukan (Hendrawan dkk, 2021) "Klasifikasi Lama Studi dan Predikat Kelulusan Mahasiswa menggunakan Metode Naïve Bayes" pada penelitian ini diklasifikasikan lama studi dan predikat kelulusan mahasiswa dengan tujuan untuk membantu pihak program studi dan fakultas dalam menganalisis luaran pembelajaran. Metode klasifikasi yang diterapkan pada penelitian ini adalah Naïve Bayes. Data yang digunakan adalah data mahasiswa Institut Teknologi dan Bisnis STIKOM Bali tahun 2008 sampai dengan tahun 2016 dengan total jumlah

data sebanyak 5.081. Atribut dataset yang digunakan untuk mengklasifikasikan lama Studi dan Predikat Kelulusan adalah Jenis Kelamin, Prodi, Konsentrasi, Tahun Masuk, dan Tahun Lulus. Hasil eksperimen menunjukkan bahwa akurasi tes classifier untuk klasifikasi lama studi sebesar 0,74 dan untuk akurasi tes klasifikasi predikat kelulusan sebesar 0,61 pada kelompok data Program Studi Sistem Komputer. Kemudian untuk kelompok data Program Studi Sistem Informasi akurasi tes klasifikasi lama studi sebesar 0,73 dan untuk akurasi klasifikasi predikat kelulusan sebesar 0,67.

Dalam penelitian ini penulis akan membandingan tiga metode data mining yaitu metode Naive Bayes Classifier, Decision Tree dan K-Nearest Neighbors. dalam pemodelan prediksi ketepatan waktu lulus mahasiswa dengan mengukur akurasi hasil prediksi menggunakan teknik scenario uji n-Folds Cross Validation.

1.2. Rumusan Masalah

Berdasarkan latar belakang penelitian diatas maka diperoleh rumusan masalah sebagai berikut:

a. Bagaimana evaluasi implementasi algoritma Naïve Bayes Classification, Decision Tree dan K-Nearest Neighbors dalam pemodelan prediksi ketepatan waktu lulus mahasiswa dengan mengukur tingkat akurasi, presisi, recall dan F1-Score menggunakan teknik skenario uji n-Folds Cross Validation.

b. Algoritma mana yang memiliki tingkat akurasi terbaik untuk kasus prediksi ketepatan waktu lulus mahasiswa tepat dan terlambat di Poltekkes Permata Indonesia Yogyakarta?

1.3. Batasan Masalah

Batasan masalah yang digunakan dalam penelitian ini adalah sebagai berkut:

a. Data yang digunakan dalam penelitian adalah data mahasiswa Poltekkes Permata Indonesia Yogyakarta untuk Tahun Akademik (TA) 2019/2020. Pertimbangan penggunaan keseluruhan data mahasiswa, baik yang sudah memiliki status akhir lulus maupun belum, dikarenakan agar sistem dapat menampilkan model prediksi yang akurat dengan tidak menghapus data sebenarnya.
b. Penentuan atribut-atribut yang mempengaruhi kelulusan mahasiswa didapatkan berdasarkan hasil studi literature dan focus group discussion yang selanjutnya divalidasi oleh ahli/decision maker dalam hal ini adalah pengelola Program Studi dan bagian adnisi atau penerimaan mahasiswa baru,
c. Penelitian ini tidak membahas masalah mengenai analisa pada kompleksitas algoritma.

1.4. Tujuan Penelitian

Tujuan dari penelitian ini adalah sebagai berikut:

a. Mengevaluasi implementasi algoritma Naïve Bayes Classification, Decision Tree dan K-Nearest Neighbors dalam pemodelan prediksi ketepatan waktu lulus mahasiswa dengan mengukur tingkat akurasi menggunakan teknik skenario uji n-Folds Cross Validation.
b. Mengetahui algoritma mana yang memiliki tingkat akurasi yang lebih tinggi untuk kasus prediksi ketepatan waktu lulus mahasiswa tepat dan terlambat di Permata Indonesia Yogyakarta.

1.5. Manfaat Penelitian

Adapun manfaat dari penelitian yang dilaksanakan penulis pada Poltekkes Permata Indonesia Yogyakarta adalah sebagai berikut:

a. Dapat menjadi bahan evaluasi bagian program studi (Prodi) Poltekkes Permata Indonesia Yogyakarta untuk mengantisipasi banyaknya mahasiswa yang lulus tidak tepat waktu;
b. Dapat menjadi acuan bagi Prodi agar dapat memberi perhatian khusus terhadap mahasiswa yang diprediksi tidak lulus tepat waktu.

BAB II

TINJAUAN PUSTAKA

2.1. Tinjauan Pustaka

Penelitian terdahulu yang pernah dilakukan, relevan dan dijadikan studi literatur adalah sebagai berikut: Pada penelitian (Nabila dkk, 2021) yang berjudul "Model Prediksi Kelulusan Tepat Waktu Dengan Metode Fuzzy C-Means Dan K-Nearest Neighbors Menggunakan Data Registrasi

Mahasiswa" Penelitian ini merupakan penggabungan antara algoritma Fuzzy CMeans dan K-Nearest Neighbors, menggunakan bahasa pemrograman python dengan tools Jupyter Notebook, dataset yang digunakan data registrasi mahasiswa UINSA, kemudian pengujian skor akan digunakan confusion matrix dan k-fold cross validation. Hasil dari algoritma FCM-KNN didapatkan bahwa model prediksi dengan pengujian 10-fold cross validation dengan skenario k=1 mempunyai rata rata akurasi sebesar 71% .(Nabila dkk, 2021)

Penelitian yang dilakukan (Mulia dan Muanas, 2021) "Model Prediksi Kelulusan Mahasiswa Menggunakan Decision Tree C4.5 dan Software Weka" Dalam penelitian ini dibangun sebuah model untuk memprediksi status kelulusan mahasiswa Institut Bisnis dan Informatika Kesatuan menggunakan algoritma pohon keputusan C4.5. Model prediksi dibangun dengan menggunakan IPK mahasiswa semester 1 sampai semester 4, untuk mahasiswa tahun masuk 2013 sampai 2016. Model prediksi yang didapat adalah pohon keputusan dengan 26 aturan, dengan atribut IPS_4 menjadi atribut yang

menentukan label kelulusan dari siswa. Model prediksi ini menghasilkan akurasi sebesar 73%, hasil yang kurang baik. Hasil ini kemungkinan disebabkan oleh proporsi data yang digunakan tidak seimbang.

Penelitian lain (Hendrawan dkk, 2021) "Klasifikasi Lama Studi dan Predikat Kelulusan Mahasiswa menggunakan Metode Naïve Bayes" pada

penelitian ini diklasifikasikan lama studi dan predikat kelulusan mahasiswa dengan tujuan untuk membantu pihak program studi dan fakultas dalam menganalisis luaran pembelajaran. Metode klasifikasi yang diterapkan pada penelitian ini adalah

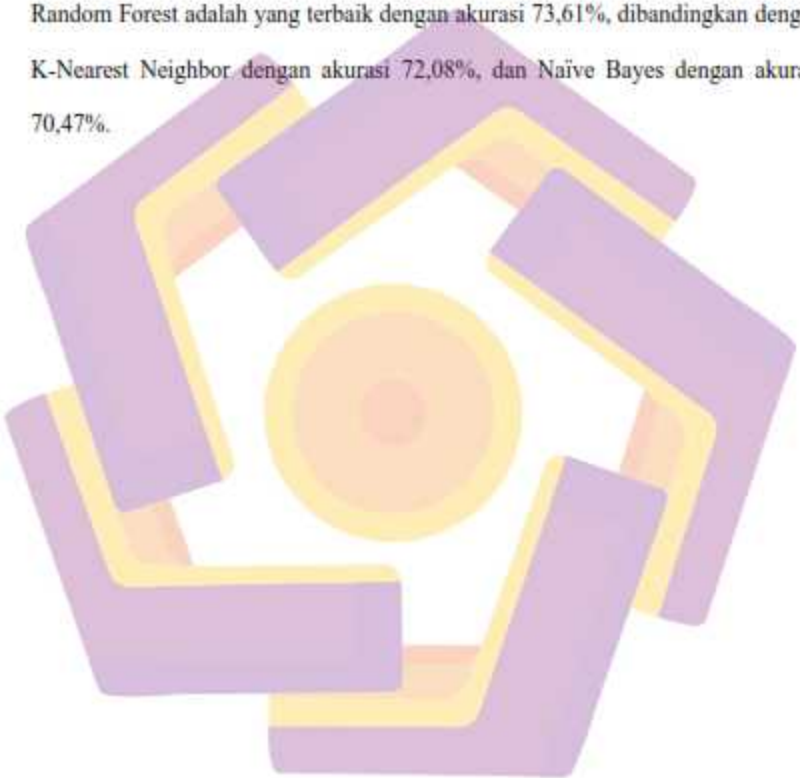
Naïve Bayes. Data yang digunakan adalah data mahasiswa Institut Teknologi dan Bisnis STIKOM Bali tahun 2008 sampai dengan tahun 2016 dengan total jumlah data sebanyak 5.081. Atribut dataset yang digunakan untuk mengklasifikasikan Lama Studi dan Predikat Kelulusan adalah Jenis Kelamin, Prodi, Konsentrasi, Tahun Masuk, dan Tahun Lulus. Hasil eksperimen menunjukkan bahwa akurasi tes classifier untuk klasifikasi lama studi sebesar 0,74 dan untuk akurasi tes klasifikasi predikat kelulusan sebesar 0,61 pada kelompok data Program Studi Sistem Komputer. Kemudian untuk kelompok data Program Studi Sistem Informasi akurasi tes klasifikasi lama studi sebesar 0,73 dan untuk akurasi klasifikasi predikat kelulusan sebesar 0,67 Penelitian (Farhana, 2021) "Classification of Academic Performance for University Research Evaluation by Implementing Modified Naive Bayes Algorithm" Karya penelitian ini menyajikan klasifikasi awal dan prediksi penelitian dan kinerja akademik universitas staf

sebelum mengevaluasi Research Assessment (RA) dengan mengimplementasikan algoritma klasifikasi Modified Naïve Bayes. Di dalam klasifikasi manual, kinerja peneliti dan staf akademik diamati oleh penanggung jawab yang terkait dengan penelitian dan merekam data. Dalam klasifikasi eksperimental yang mengikuti proses ini, pertama, model dihasilkan menggunakan kumpulan data pelatihan, selanjutnya model diuji dengan testing dataset tanpa kelas atribut. Sebagai hasil dari penelitian tersebut, klasifikasi Modified Naïve Bayes dapat mengklasifikasikan kinerja akademik dengan kegiatan penelitian lebih baik daripada pohon keputusan dan Naïve Bayes biasa, masingmasing 96,15% dan 94,23%.

Penelitian "Analisis Komparasi Algoritma Naïve Bayes dan K-Nearest Neighbor Untuk Memprediksi Kelulusan Mahasiswa Tepat Waktu" Penelitian ini akan mencoba untuk membandingkan hasil Analisa dua metode dalam algoritma klasifikasi untuk memprediksi kelulusan mahasiswa. Algoritma yang digunakan ialah Algoritma K-Nearest Neighbour dan Naïve Bayes. Penelitian ini menghasilkan kesimpulan bahwa algoritma Naïve Bayes memiliki tingkat akurasi yang sama dengan algoritma KNN dalam memprediksi kelulusan mahasiswa program studi Pendidikan Kedokteran yaitu sebesar 90 %.(Gunawan dkk, 2021)

Penelitian (Sejati dkk, 2019) yang berjudul "Studi Komparasi Naive Bayes, K-Nearest Neighbor, Dan Random Forest Untuk Prediksi Calon Mahasiswa Yang Diterima Atau Mundur" Penelitian ini bertujuan untuk mendapatkan model prediksi terbaik dari data Penerimaan Mahasiswa Baru tahun 2014 hingga 2019

dengan membandingkan Naive Bayes, K-Nearest Neighbor, dan Random Forest. Penelitian ini menggunakan metode klasifikasi untuk memprediksi calon mahasiswa. Mereka diterima atau mundur. Dalam penelitian ini digunakan 19.603 data latih dan 4.901 data uji. Hasil penelitian menunjukkan bahwa algoritma Random Forest adalah yang terbaik dengan akurasi 73,61%, dibandingkan dengan K-Nearest Neighbor dengan akurasi 72,08%, dan Naïve Bayes dengan akurasi 70,47%.



2.2. Keaslian Penelitian

Tabel 2.1. Matriks literatur review dan posisi penelitian

	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
1	Studi Komparasi Naïve Bayes, K-Nearest Neighbor, Dan Random Forest Untuk Prediksi Calon Mahasiswa Yang Diterima Atau Mundur	Puteri Sejati, Munawar, Marzuki Pilliang, Habibullah Akbar, JTIIK 2019 (Sejati dkk, 2019)	Prediksi Calon Mahasiswa Yang Diterima Atau Mundur	Hasil penelitian menunjukkan bahwa algoritma Random Forest adalah yang terbaik dengan akurasi 73,61%,dibandingkan dengan K-Nearest Neighbor dengan akurasi 72,08%, dan Naive Bayes dengan akurasi 70,47%.	Pada proses optimasi di Random Forest memakan waktu yang lama. Sehingga masih diperlukan penelitian lanjutan agar mendapatkan model optimal dengan waktu yang relatif singkat.	Pada penelitian sebelumnya tidak menentukan atribut yang paling berpengaruh dalam rekomendasi calon mahasiswa baru sedang dalam penelitian ini akan menentukan atribut yang paling dominaan dalam memprediksi kelulusan mahasiswa.
2	Classification of Academic Performance for University Research Evaluation by Implementing Modified Naïve	Soheli Farhana, Elsavier, 2021	Menganalisis dengan membandingkan kedua teknik Evaluation by Implementing Modified Naïve Bayes	klasifikasi Modified Naïve Bayes Dapat mengklasifikasikan kinerja akademik dengan kegiatan penelitian lebih baik daripada Naïve Bayes biasa,94,23 %	Bahwa atribut yang diidentifikasi sebelum awal data hibah penelitian relevan contributor penilaian RA. Informasi ini dapat bermanfaat bagi spesialis yang tertarik untuk	Pada penelitian sebelumnya decision tree mendapatkan hasil yang lebih baik dari pada Naïve Bayes

Tabel 2.1 Tabel Lanjutan

	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
	Bayes Algorithm		Algorithm	Decision tree 96,15%	memahami kemampuan peneliti untuk mengakses kinerja akademik.	
3	Sistem Prediksi Lama Studi Kuliah Menggunakan Metode Naive Bayes	Muqorobin, Moch Bagoes Pakarti, JIKOBIS, 2021	Prediksi Lama Studi Kuliah	Berdasarkan proses perhitungan uji model dengan menggunakan metode Confusion Matrik pada 13 data training dan 5 data testing telah diperoleh hasil Akurasi 80%, recall 50% dan presisi 100%	Maka bagi peneliti selanjutnya dapat mengembangkan penelitini degnan membuat aplikasi sistem prediksi sehingga hal tersebut dapat memberikan kontribusi lebih bagi kampus dan perguruan teinggi yang lainnya.	Pada penelitian sebelumnya tidak menentukan atribut yang paling berpengaruh dalam rekomendasi calon mahasiswa baru sedang dalam penelitian ini akan menentukan atribut yang paling dominaan dalam memprediksi kelulusan mahasiswa.
4	Klasifikasi Lama Studi dan Predikat Kelulusan Mahasiswa menggunakan Metode Naïve Bayes	I Nyoman R Hendrawan, I Made Arya B S, Gusti Ayu Putu Cahya Dewi, I Gede Surya Adi Pranata, Ni	Klasifikasi Lama Studi dan Predikat Kelulusan Mahasiswa	Akurasi tes classifier untuk klasifikasi lama studi 0,74 dan untuk akurasi tes klasifikasi predikat kelulusan 0,61	Kinerja classifier Naïve Bayes tidak cukup baik pada kelas Predikat Kelulusan baik itu dari kelompok data Program Studi Sistem Komputer ataupun Program Studi Sistem Informasi, hal ini	Pada penelitian sebelumnya tidak menentukan atribut yang paling berpengaruh dalam rekomendasi calon mahasiswa baru sedang dalam penelitian ini akan menentukan atribut yang paling dominaan dalam

Tabel 2.1 Tabel Lanjutan

	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
		Luh Nyoman Wedasari, Jurnal Eksplora Informatika, 2021			dikarenakan adanya imbalanceddata.	memprediksi kelulusan mahasiswa.
5	Model Prediksi Kelulusan Mahasiswa Menggunakan Decision Tree C4.5 dan Software Weka	Isnan Mulia, Muanas, 2021	Prediksi Kelulusan Mahasiswa Menggunakan Decision Tree C4.5	Model prediksi ini menghasilkan akurasi sebesar 73%, hasil yang kurang baik. Hasil ini kemungkinan disebabkan oleh proporsi data yang digunakan tidak seimbang	pada penelitian ini memiliki hasil yang kurang baik. Saran bisa menggunakan metode yang lain, misalnya KNearest Neighbor, Neural Network, atau Support Vector Machine.	Pada penelitian sebelumnya tidak menentukan atribut yang paling berpengaruh dalam prediksi kelulusan mahasiswa
...	Algoritma Decision Tree C.45 Dalam Analisa Kelulusan Mahasiswa Program Studi	Aslam Fatkhudin, M. Yusuf Febrianto, Fenilinas Adi Artanto, M. Waffa	Analisa Kelulusan Mahasiswa Program Studi Manajemen Informatika	Dari analisa yang dilakukan didapatkan bahwa tahun masuk mahasiswa menjadi faktor variabel utama dalam kelulusan mahasiswa dengan	Variabel faktor penunjang kelulusan mahasiswa belum banyak dan penggunaan metode klasifikasi hanya sebatas decesion tree saja. Bisa	Pada penelitian sebelumnya tidak menentukan atribut yang paling berpengaruh dalam prediksi kelulusan mahasiswa sedang dalam penelitian ini akan menentukan atribut yang

Tabel 2.1 Tabel Lanjutan

	Judul	Peneliti, Media Publikasi, dan Tahun	Tujuan Penelitian	Kesimpulan	Saran atau Kelemahan	Perbandingan
	Manajemen Informatika Umpp	Najib Hadinata, Riza Fahlevi, Jurnal Ilmiah Ilmu Komputer, 2021		didapatkan akurasi algoritma decision tree sebesar 73,48%.	menggunakan metode lainya yang juga terdapat penambahan jumlah data dan jumlah variabelny	paling dominaan dalam memprediksi kelulusan mahasiswa.

2.3. Landasan Teori

2.3.1 Data Mining

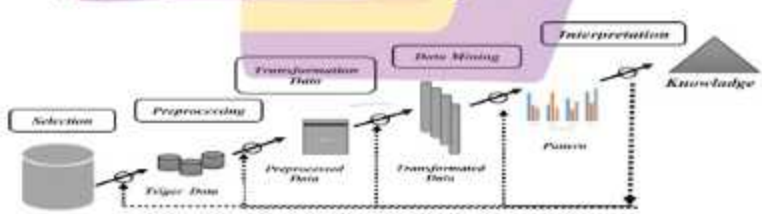
Data mining merupakan salah satu teknik untuk menggali atau "menambang" pengetahuan dari sekumpulan besar data. Data mining merupakan analisis dari peninjauan kumpulan data untuk menemukan hubungan yang tidak diduga dan meringkas data dengan cara yang berbeda dengan sebelumnya yang dapat dipahami dan bermanfaat bagi pemilik data (Larose, 2005). Terdapat beberapa teknik yang digunakan untuk data mining seperti yang diungkapkan Turban, et al (2011) data mining adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan di dalam database. Data mining adalah proses yang menggunakanteknik statistik, matematika, kecerdasan buatan dan machine learning untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar.

Data mining biasanya mengolah data dari database dengan ukuran yang besar. Dari data tersebut dilakukan pencarian pola atau trend sesuai dengan tujuan dari penerapan data mining tersebut. Hasil dari pengolahan data mining tersebut selanjutnya dapat digunakan untuk pengambilan keputusan maupun analisis yang dibutuhkan. Terdapat beberapa alasan mengapa ilmu data mining dibutuhkan saat ini diantaranya terdapat sejumlah besar data di suatu perusahaan atau organisasi yang hanya tersimpan di dalam database tanpa dianalisis lebih lanjut untuk digunakan untuk perkembangan perusahaan atau organisasi tersebut. Selain itu dengan perkembangan internet yang sangat pesat, memberikan dampak positif

dengan kemudahan akses data dengan berbagai perangkat hardware dan software yang memiliki daya komputasi dan kapasitas yang luar biasa. Sedangkan dilihat lingkungan luar, tekanan kompetisi untuk memperluas pangsa pasar dan keuntungan juga semakin meningkat sehingga dibutuhkan cara lain dengan menggali informasi yang tersimpan pada data yang dimiliki perusahaan atau organisasi tersebut. Meskipun algoritma data mining biasanya diterapkan untuk ukuran data yang besar, beberapa algoritma bisa juga diterapkan untuk ukuran data.

2.3.2 Knowledge Discovery in Databases (KDD)

Knowledge Discovery in Databases (KDD) adalah keseluruhan proses untuk mengkonversi data mentah menjadi suatu pengetahuan yang bermanfaat. Istilah data mining dan Knowledge Discovery in Databases (KDD) sering kali digunakan secara bergantian untuk menjelaskan proses penggalian informasi tersembunyi dalam suatu basis data yang besar. Sebenarnya kedua istilah tersebut memiliki konsep yang berbeda, tetapi berkaitan satu sama lain. Salah satu tahapan dalam keseluruhan proses KDD adalah data mining. Dapat di lihat Tahapan Proses KDD yang di tunjukkan pada Gambar 2.1. (Fayyad, 1996).



Gambar 2. 1 Tahapan Proses KDD

Tahapan proses KDD dapat di jelaskan sebagai berikut :

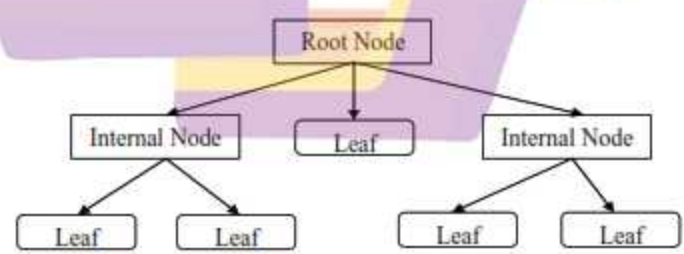
1. Data Selection Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data hasil seleksi yang akan digunakan, disimpan dalam suatu berkas, terpisah dari basis data operasional.
2. Pre-processing/Cleaning Sebelum proses data mining, perlu dilakukan proses cleaning pada data yang menjadi fokus KDD. Proses cleaning mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak.
3. Transformation Coding adalah proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses data mining. Proses coding dalam KDD merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data.
4. Data mining Data mining adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik tertentu. Teknik, metode, atau algoritma dalam data mining sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.
5. Interpretation Evaluation Pola informasi yang dihasilkan dari proses data mining perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan. Tahap ini merupakan bagian dari proses KDD yang disebut interpretation. Tahap ini mencakup pemeriksaan apakah pola atau

informasi yang ditemukan bertentangan dengan fakta atau hipotesis yang ada sebelumnya. Tahapan dari KDD dapat dilihat pada Gambar 1.

Model Decision Tree *Decision tree* adalah *flow-chart* seperti *struktur tree*, dimana tiap *internal node* menunjukkan sebuah test pada sebuah atribut, tiap cabang menunjukkan hasil dari test, dan *leaf node* menunjukkan *class-class* atau *class distribution*. Selain karena pembangunannya relatif cepat, hasil dari model yang dibangun mudah untuk dipahami. Pada *decision tree* terdapat 3 jenis *node*, yaitu:

1. *Root Node*, merupakan *node* paling atas, pada *node* ini tidak ada *input* dan bisa tidak mempunyai *output* atau mempunyai *output* lebih dari satu.
2. *Internal Node*, merupakan *node* percabangan, pada *node* ini hanya terdapat satu *input* dan mempunyai *output* minimal dua.
3. *Leaf node* atau *terminal node*, merupakan *node* akhir, pada *node* ini hanya terdapat satu *input* dan tidak mempunyai *output*.

Contoh dari model pohon keputusan di tunjukkan pada gambar 2.2



Gambar 2. 2 Model Decision Tree

2.3.3 Decision Tree C4.5

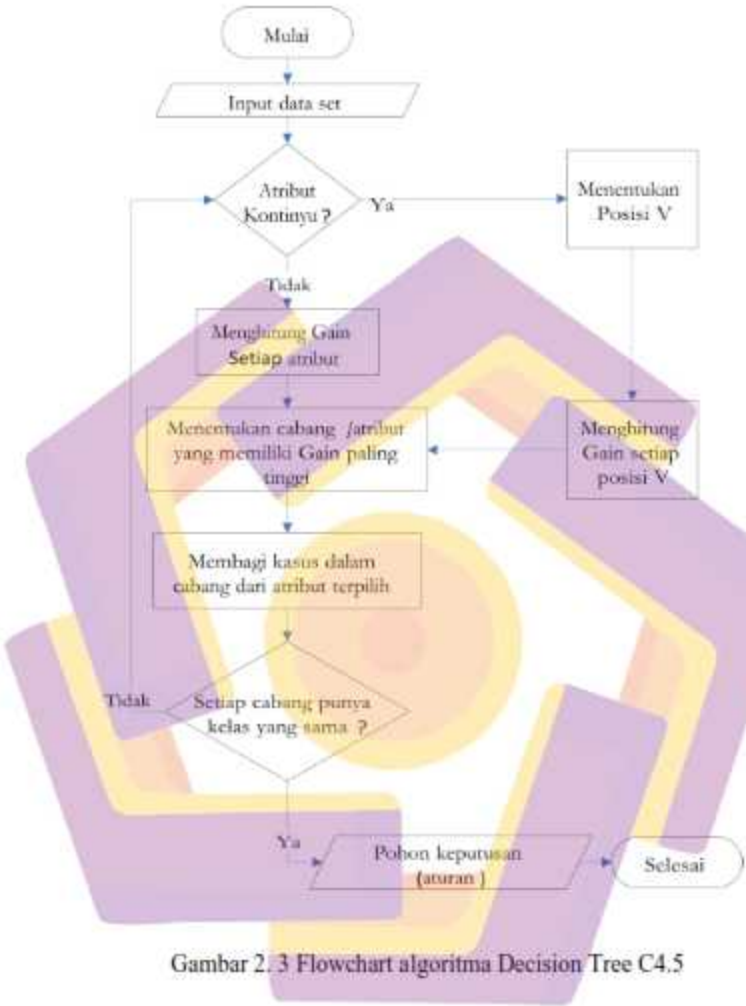
C4.5 adalah algoritma yang sudah banyak dikenal dan digunakan untuk klasifikasi data yang memiliki atribut-atribut numerik dan kategorikal. Hasil dari proses klasifikasi yang berupa aturan-aturan dapat digunakan untuk memprediksi nilai atribut bertipe *diskret* dari *record* yang baru.

Algoritma C4.5 sendiri merupakan pengembangan dari algoritma ID3, dimana pengembangan dilakukan dalam hal bisa mengatasi *missing* data, bisa mengatasi data *kontinyu*, dan *pruning*.

Secara umum algoritma C4.5 untuk membangun pohon keputusan adalah sebagai berikut:

1. Pilih atribut sebagai akar.
2. Buat cabang untuk tiap-tiap nilai.
3. Bagi kasus dalam cabang.
4. Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama.

Berikut ini akan dijelaskan secara lebih detail algoritma C4.5 menggunakan *flowcart* yang di tunjukkan pada gambar 2.3.



Gambar 2. 3 Flowchart algoritma Decision Tree C4.5

Untuk memilih atribut sebagai simpul akar (*root node*) atau simpul dalam (*internal node*), didasarkan pada nilai *information gain* tertinggi dari atributatribut yang ada. Sebelum perhitungan *information gain*, akan dilakukan perhitungan *entropy*. *Entropy* merupakan distribusi probabilitas dalam teori informasi dan diadopsi kedalam algoritma *C*4.5 untuk

mengukur tingkat homogenitas distribusi kelas dari sebuah himpunan data (*data set*). Semakin tinggi tingkat *entropy* dari sebuah data maka semakin homogen distribusi kelas pada data tersebut. Perhitungan *information gain* menggunakan rumus 1.1. sedangkan *entropy* menggunakan rumus 2.2.

$$Gain(S, A) = Entropy\ (S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy\ (S_i) \quad (2.1)$$

dimana,

S : Himpunan kasus A :

Atribut n : Jumlah partisi atribut A

|Si| : Jumlah kasus pada partisi ke i

|S| : Jumlah kasus dalam S

$$Entropy(S) = -\sum_{i=1}^{n} p_i * log_2\ p_i \quad (2.2)$$

dimana,

S : Himpunan kasus

A : Fitur

n : Jumlah partisi S

p_i : Proporsi dari S_i terhadap S

Selain *Information Gain* kriteria yang lain untuk memilih atribut sebagai pemecah adalah *Rasio Gain*. Perhitungan rasio gain menggunakan rumus 2.3, sedangkan split information menggunakan rumus 2.4.

$$GainRasio\ (S,A) = \frac{Gain\ \ (S,A)}{SplitInformation\ \ (S,A)} \quad (2.3)$$

$$SplitInformation(S,A) = -\sum_{i=1}^{c} \frac{S_i}{S}\ log_2 \frac{S_i}{S} \quad (2.4)$$

dimana S_1 sampai S_c adalah c
subset yang dihasilkan dari pemecahan S dengan menggunakan atribut A yang mempunyai sebanyak c nilai.

2.3.4 *Klasifikasi*

Algoritma data mining dapat dibagi menjadi tiga (Neelamegam & Ramaraj, 2013), yaitu supervised, unsupervised, dan semi-supervised. Dalam supervised learning, algoritma bekerja pada sekumpulan data yang telah diberi label atau telah diketahui kelasnya. Pada supervised learning, data belum diketahui label atau kelasnya, algoritma digunakan untuk mengelompokkan data berdasarkan kemiripannya. Sedangkan dalam semi supervised learning, sebagian kecil data telah memiliki label bersama dengan sejumlah data yang belum memiliki label. Klasifikasi termasuk ke dalam supervised learning. Klasifikasi dokumen adalah pemberian kategori yang telah didefinisikan kepada dokumen yang belum memiliki kategori (Goller, 2000).

Mengklasifikasi dokumen merupakan salah satu cara untuk mengorganisasikan dokumen. Dokumen-dokumen yang memiliki isi yang

sama akan dikelompokkan ke dalam kategori yang sama. Dengan demikian, orang-orang yang melakukan pencarian informasi dapat dengan mudah melewatkan kategori yang tidak relevan dengan informasi yang dicari atau yang tidak menarik perhatian (Feldman, 2004).

Pada penelitian ini, klasifikasi diterapkan untuk mengkategorikan data mahasiswa yang lulus tepat waktu dan lulus tidak tepat waktu. Klasifikasi adalah proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data dengan tujuan untuk memperkirakan kelas yang tidak diketahui dari suatu objek. Adapun dalam pengklasifikasian data terdapat dua proses yang dilakukan (Annasaheb & Verma, 2016) yaitu:

a. Tahap membangun model

   Pada langkah ini model klasifikasi dibangun berdasarkan data yang telah ditentukan kelasnya. Data sampel yang digunakan disebut sebagai data pelatihan atau data pembelajaran (training set). Proses ini disebut sebagai proses induksi. Pada proses training digunakan training set yang telah diketahui label-labelnya untuk membangun model atau fungsi.

b. Tahap menggunakan model klasifikasi

   Pada tahap ini model diterapkan pada data yang belum diketahui kelasnya. Proses penerapan model klasifikasi untuk memprediksikan kelas label dari data dalam himpunan

menggunakan data uji (testing set), proses ini disebut deduksi. Proses Testing untuk mengetahui keakuratan model atau fungsi yang akan dibangun pada proses training, maka digunakan data yang disebut dengan testing set untuk memprediksi label-labelnya

2.3.5 *Naïve Bayes Classification*

Salah satu metode yang sangat penting dalam klasifikasi adalah metode Naïve Bayes. Metode ini juga disebut idiot's Bayes, simple Bayes, independence Bayes. Yang menjadikan metode ini sangat penting karena metode ini sangat mudah dibangun, dan tidak memerlukan skema estimasi parameter berulang yang rumit. Hal ini menunjukkan bahwa metode Naïve Bayes dapat diterapkan dalam data set yang besar.

Selain itu metode Naïve Bayes sangat mudah digunakan sehingga pengguna yang tidak terampil dalam teknik klasifikasi (Wu & Kumar, 2009). Klasifikasi Bayes didasarkan pada teorema Bayes, diambil dari nama seorang ahli matematika yang juga menteri Prebysterian Inggris, Thomas Bayes (1702-1761), yaitu (Bramer, 2007) dimana teorema Bayes ditulis dengan rumus sebagai berikut:

$$P(x|y) = \frac{P(y|x)\,P(x)}{P(y)} \qquad (2.5)$$

Keterangan :

y = data dengan kelas yang belum diketahui

x = hipotesis data y merupakan suatu kelas spesifik

P(x | y) = probabilitas hipotesis x berdasar kondisi y (*posteriori probability*)

P(x) = probabilitas hipotesis x (*prior probability*)

P(y | x)= probabilitas y berdasarkan kondisi pada hipotesis x

P(y) = probabilitas dari y

Rumus teorema Bayes disederhanakan menjadi *Naïve Bayes* yang merupakan penyederhanaan metode Bayes. Penyederhanaan teorema Bayes ditulis sebagai berikut:

$$P(x|y) \sim P(y|x)\, P(x) \qquad (2.6)$$

Contoh klasifikasi dengan algoritma *Naive Bayes* yaitu dengan menggunakan data *training*. Dari data tersebut terdapat dua kelas dari klasifikasi yang dibentuk, yaitu:

C1 = *credit risk = Good*

C2 = *credit risk = Bad*

Misalkan terdapat data y (belum diketahui kelasnya) y = (*Saving = high, asset = low, income* = 50)

Penyelesaian:

Dibutuhkan P(Ci) untuk memaksimalkan P(y|Ci) untuk I = 1.2

P(Ci) merupakan *prior probability* untuk setiap class berdasarkan data contoh:

Jumlah data = 8

Jumlah data *credit risk = Good* = 5

Jumlah data *credit risk = Bad* = 3

P(cr = *good*)	= 5/8 = 0.625
P(cr = *bad*)	= 3/8 = 0.375
P(*saving* = *high* | cr = *good*)	= 1/2 = 0.5
P(*saving* = *high* | cr = *bad*)	= 1/2 = 0.5
P(*asset* = *low* | cr = *good*)	= 0
P(*asset* = *low* | cr = *bad*)	= 2/2 = 1
P(*income* <= 50 | cr = *good*)	= 2/5 = 0.4
P(*income* <= 50 | cr = *bad*)	= 3/5 = 0.6

P(X | cr = *good*)	= 0.5 · 0 · 0.4	= 0
P(X | cr = *bad*)	= 0.5 · 0.4 · 0.6 = 0.3	
P(X | cr = *good*) P(cr = *good*)	= 0 · 0.625 = 0	
P(X | cr = *bad*) P(cr = *bad*)	= 0.3 · 0.375 = 0.1125	

Dari hasil perhitungan di atas, dapat disimpulkan bahwa, data baru termasuk klasifikasi *bad* untuk *credit risk* karena nilai probabilitas *bad* lebih tinggi daripada nilai probabilitas *good*.

2.3.6 *K-Nearest Neighbors*

Algoritma *K-Nearest Neighbors* salah satu teknik klarisifikasi data yang kuat, dengan cara mencari kasus dengan menghitung kedekatan antara kasus baru dengan kasus lama berdasarkan pencocokan bobot. *K-Nearest Neighbors* adalah suatu metode algoritma supervised learning, dimana kelas yang paling banyak muncul (mayoritas) yang akan menjadi kelas hasil klasifikasi.

KNearest Neighbors merupakan contoh algoritma berbasis pembelajaran, dimana data set pelatihan (training) disimpan, sehingga klasifikasi untuk record baru yang tidak diklasifikasi didapatkan dengan membandingkan record yang paling mirip dengan training set.

Berikut adalah langkah-langkah *K-Nearest Neighbors*:

1. Menentukan parameter k (jumlah tetangga paling dekat), Parameter k pada testing ditentukan berdasarkan nilai k optimum pada saat training.
2. Menghitung kuadrat jarak euclid (euclidean distance) masing-masing objek terhadap data sampel yang diberikan.
3. Mengurutkan objek-objek tersebut kedalam kelompok yang mempunyai jarak Euclidian terkecil.
4. Mengumpulkan kategori Y (*klasifikasi Nearest Neighbors*).

Dengan menggunakan kategori mayoritas, maka dapat hasil klasifikasi Secara umum untuk mendefinisikan jarak antara dua objek x dan y, digunakan rumus jarak Euclidian pada persamaan:

$$d(x, y) = \sqrt{\sum n \; i=1(xi - yi)2} \quad \text{.....(2.7)}$$

Keterangan:

x : data training ke-i,

y : data testing

n : jumlah data training.

Record (baris) ke-i dari tabel, Dimana matriks distance adalah jarak skala dari kedua vektor x dan y dari matriks dengan ukuran dimensi. Pada fase training, algoritma ini hanya melakukan penyimpanan vektor-vektor fitur dan klasifikasi data training sample. Pada fase klasifikasi, fitur-fitur yang sama dihitung untuk testing data (yang klasifikasinya tidak diketahui). Jarak dari vektor baru yang ini terhadap seluruh vektor training sample dihitung dan sejumlah k buah yang paling dekat diambil.

2.3.7 *Evaluasi Performansi Metode Klasifikasi*

Evaluasi dan validasi hasil klasifikasi dengan data mining pada penelitian ini digunakan metode *Confusion Matrix* dan kurva ROC (*Receiver Operating Characteristic*).

1. Confusion Matrix

Metode ini hanya menggunakan tabel matriks seperti pada Tabel 2.1, jika dataset hanya terdiri dari dua kelas, kelas yang satu dianggap sebagai positif dan yang lainnya negatif (Bramer, 2007).

Evaluasi dengan *confusion matrix* menghasilkan nilai *accuracy*, *precison*, dan *recall*. *Accuracy* dalam klasifikasi adalah persentase ketepatan *record* data yang diklasifikasikan secara benar setelah dilakukan pengujian pada hasil klasifikasi (Han & Kamber, 2006). Sedangkan *precision* atau *confidence* adalah proporsi kasus yang diprediksi positif yang juga positif benar pada data yang sebenarnya. *Recall* atau *sensitivity* adalah proporsi kasus positif yang sebenarnya yang diprediksi positif secara benar (Powers, 2011).

Confusion Matrix Tebentuk berdasarkan empat hasil klasifikasi biner (Hussain, 2018). Dalam klasifikasi biner, biasanya dataset memiliki dua label positif (P) dan negative (N). Hasilnya berupa True Positif (TP) yaitu prediksi positif yang benar, True Negatif (TN) yaitu prediksi negative yang benar, False Positif (FP) yaitu prediksi positif yang salah dan False Negatif (FN) yaitu prediksi negative yang salah. Permasalahan pada klasifikasi biner, akurasi klasifikasi ditunjukkan pada Tabel 2.1.

Tabel 2. 1 Akurasi Klasifikasi

Aktual	Prediksi	
	Positif	Negatif
Positif	TP	FN
Negatif	FP	TN

Keterangan:

TP: *True Positive* (Jumlah prediksi benar pada kelas positif)

FP: *False Positive* (Jumlah prediksi salah pada kelas positif)

FN: *False Negative* (Jumlah prediksi salah pada kelas negatif)

TN: *True Negative* (Jumlah prediksi benar pada kelas negatif)

Pada Tabel 2.1 nilai TP (*true positive*) dan TN (*true negative*) menunjukkan tingkat ketepatan klasifikasi. Umumnya semakin tinggi nilai TP dan TN semakin baik pula tingkat klasifikasi dari akurasi, presisi, dan *recall*. Jika label prediksi keluaran bernilai benar (true) dan nilai sebenarnya bernilai salah (false) disebut sebagai *false positive* (FP). Sedangkan jika prediksi label keluaran bernilai salah (false) dan

nilai sebenarnya bernilai benar (true) maka hal ini disebut sebagai *false negative* (FN). Perhitungan atau rumus dari pengukuran kinerja menggunakan *confusion matrix* meliputi: Recall (persamaan 2.8), Presisi (persamaan 2.9), dan Akurasi (persamaan 2.10)

a. Sensitivity (Recall or True positive rate)

Sensitivitas (*Recall*) adalah jumlah klasifikasi yang benar dibagi dengan jumlah total positif. Jadi,

$$Recall = TP/(TP + FN) + TP/P \quad (2.8)$$

b. Presisi (*Precision*)

Presisi (*Precision*) adalah jumlah klasifikasi positif yang benar dibagi dengan jumlah total klasifikasi positif. Jadi,

$$Precision = TP/(TP+FP) \quad (2.9)$$

c. Akurasi (*Accuracy*)

Akurasi (*Accuracy*) adalah jumlah semua klasifikasi yang benar dibagi dengan jumlah kasus. Jadi,

$$Accuracy = (TP + TN)/(TP + TN + FN + FP) = (TP + TN)/(P + N) \quad (2.10)$$

2. Kurva ROC

Kurva ROC menunjukkan akurasi dan membandingkan klasifikasi secara visual. ROC mengekspresikan *confusion matrix*. ROC adalah grafik dua dimensi dengan *false positives* sebagai garis horizontal dan *true positive* sebagai garis vertical (Vercellis, 2009). *The area under curve* (AUC) dihitung untuk mengukur perbedaan performansi metode yang digunakan. AUC digunakan dengan menggunakan rumus (Liao, 2007):

$$\theta^r = \frac{1}{mn} \sum_{j=1}^{n} \sum_{i=1}^{m} \psi\,(xt^r, xj^r) \qquad (2.11)$$

Dimana :

$$\psi(X,Y) = \begin{cases} 1 & Y < X \\ \frac{1}{2} & Y = X \\ 0 & Y > X \end{cases} \qquad (2.12)$$

Keterangan :

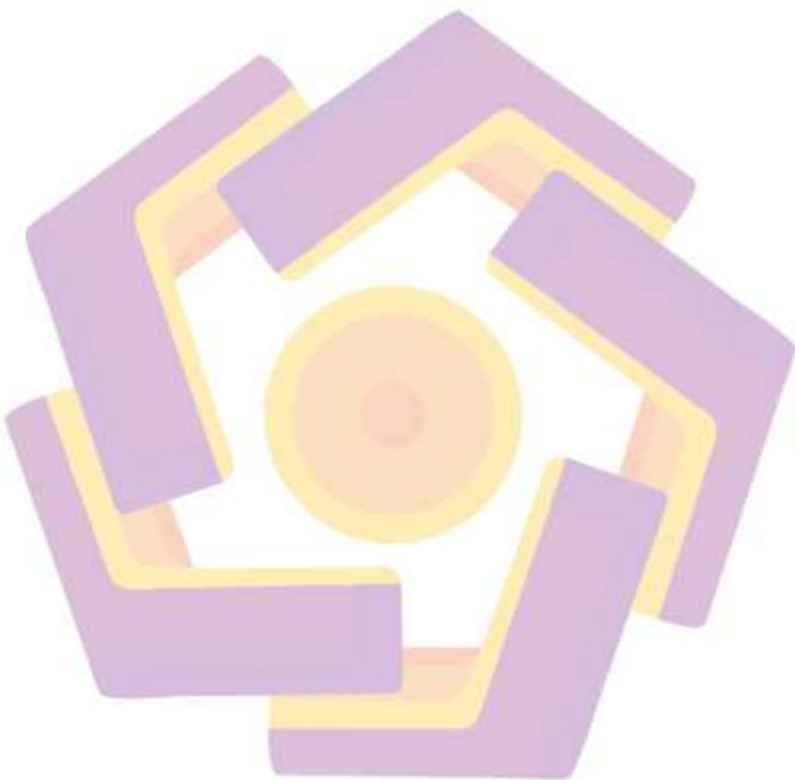
X = Output positif

Y = Output negatif

2.3.8 *Cross Validation*

Cross validation merupakan salah satu teknik untuk menilai atau validasi keakuratan sebuah model berdasarkan dataset tertentu. Dalam pengujian menggunakan K-fold cross validation disebut data training, sedangkan data yang digunakan untuk validasi model disebut data testing. Dataset dibagi menjadi sejumlah K-Fold secara acak. Kemudian dilakukan sejumlah K-kali eksperimen, dimana setiap eksperimen menggunakan data partisi ke-K sebagai data testing dan memanfaatkan sisa partisi lainnya sebagai data training. Proses ini diulangi sebanyak k subsets dan hasil akurasi klasifikasi yaitu hasil

rata-rata dari setiap data training dan testing. K-Folds yang biasa digunakan adalah 3, 5, 10 dan 20 (Bolon, Sanchez & Alonso, 2015).



BAB III

METODE PENELITIAN

3.1. Jenis, Sifat, dan Pendekatan Penelitian

Dalam penelitian ini mengunakan jenis penelitian eksperimen, yaitu dilakukan analisis komparasi menggunakan tiga metode klasifikasi pengujian performance dan memprediksi menggunkan algoritma C4.5, Naïve Bayes dan KNearest Neighbors. Dalam prediksi kelulusan mahasiswa tepat waktu mengunkan data kelulusan dan data mahasiswa di Poltekkes Permata Indonesia.

Adapun penelitian ini dilakukan secara mandiri mengunkan sifat deskriptif dimana dari data yang di peroleh kemudian dibandingakan dengan data yang diuji. Pendekatan penelitian ini menggunakan pendekatan kuantitatif.

Penelitian ini menggunakan data kelulusan dan data mahasiswa Poltekkes Permata Indonesia Yogyakarta. Data eksperimen diambil dari data kelulusan dan data mahasiswa tahun 2019 kemudian diolah dan dihitung menggunkan algoritma *C4.5, Naïve Bayes dan K-Nearest Neighbors*.

3.2. Metode Pengumpulan Data

Dalam pengumpulan data, peneliti mengambil data kelulusan mahasiswa dari bagian kemahasiswaan dan alumni Poltekkes Permata Indonesia Yogyakarta. Adapun sumber data yang digunakan dalam penelitian ini yaitu data primer dan data sekunder. Data primer merupakan data yang di dapat langsung dari bagian

akademik Poltekkes Permata Indonesia Yogyakarta, sedangkan data sekunder yaitu sumber data yang diperoleh dari media perantara atau secara tidak langsung yaitu berupa buku dan jurnal. Untuk mencari data sekunder peneliti akan melakukan pencarian di internet dan melakukan kunjungan ke perpustakaan.

Data yang diperolah adalah data sekunder karena diperoleh dari database mahasiswa yang dimiliki oleh Poltekkes Permata Indonesia yang berada di Yogyakarta, yaitu melalui Bagian Akademik. Data yang diperoleh dalam penelitian ini adalah data kualitatif dan kuantitatif. Data yang dikumpulkan adalah data mahasiswa Poltekkes Permata Indonesia dengan program studi Diploma tiga (D3) untuk tahun kelulusan periode september 2022. Data terkumpul sebanyak 305 data, dengan atribut nim, nama, umur, Program studi, IP semester 1, IP semester 2, IP semester 3 sampai dengan IP Semester 6, dengan label terlambat dan tepat. Sampel data yang diperoleh untuk penelitian dapat dilihat pada tabel 3.1.

Tabel 3. 1 Data Mahasiswa

No	NIM	Nama Mahasiswa	Program Studi	Jenis Kelamin	Umur	IPS1	IPS2	IPS3	IPS4
1	2019.131.001	Alfin Nurul Mawadda	D3 Kebidanan	P	23	4.00	3.96	3.97	3.98
2	2019.131.002	Aluysia Putri Larasati	D3 Kebidanan	P	23	3.35	3.37	3.35	3.43
2	2019.131.003	Angelia Protestia Ningrum Ate	D3 Kebidanan	P	25	3.17	3.09	3.00	3.05
3	2019.131.004	Agreni Loru	D3 Kebidanan	P	24	3.43	3.26	3.22	3.22
4	2019.131.005	Apliana Rode	D3 Kebidanan	P	23	3.04	3.02	2.94	2.95
5	2019.131.006	Aprianti Tamo Ina	D3 Kebidanan	P	22	3.09	3.20	3.26	3.32
6	2019.131.007	Arnis Tiyawika Lasoma	D3 Kebidanan	P	24	4.00	3.83	3.76	3.73

Tabel 3. 2 (Lanjutan)

7	2019.131.008	Clarizza Yublina Ayu April Mali	D3 Kebidanan	P	22	3.09	3.15	3.08	3.10
8	2019.131.009	Debi Yanti Lende	D3 Kebidanan	P	25	3.13	3.22	3.28	3.31
9	2019.131.010	Diana A Wuarbanaran	D3 Kebidanan	P	22	3.09	3.17	3.16	3.19
10	2019.131.011	Efrosina Yuan Ndelo	D3 Kebidanan	P	23	3.43	3.37	3.40	3.47
11	2019.131.012	Elyatuzzukfa Sandrias Aridita	D3 Kebidanan	P	23	4.00	0.00	0.00	0.00
12	2019.131.013	Evayati Sisiana Bili	D3 Kebidanan	P	23	3.35	3.07	3.09	3.14
13	2019.131.014	Helmania Caetrin	D3 Kebidanan	P	22	3.52	3.50	3.35	3.39
14	2019.131.016	Lortarika Juwita Ari Manna	D3 Kebidanan	P	22	3.39	3.35	3.46	3.42
15	2019.131.017	Nurhikmah	D3 Kebidanan	P	22	3.26	3.46	3.50	3.57
16	2019.131.018	Onike Rosmiati Ngongo	D3 Kebidanan	P	22	2.74	2.96	3.00	3.09
17	2019.131.019	Orasinta Landi	D3 Kebidanan	P	22	2.61	2.78	2.93	2.97
18	2019.131.020	Rurin Retno Sari	D3 Kebidanan	P	23	3.83	3.70	3.71	3.70
19	2019.131.021	Serliana Wona Haghu	D3 Kebidanan	P	22	2.52	2.78	2.72	2.76
20	2019.131.022	Yohana Mone	D3 Kebidanan	P	24	2.35	2.72	2.72	2.83
21	2019.131.023	Yuniati Ince Mone	D3 Kebidanan	P	22	2.48	2.65	2.72	2.85
22	2019.131.024	Mukmina Lorce kiha	D3 Kebidanan	P	23	3.13	3.22	3.18	3.16
23	2019.131.025	Apliana Holo	D3 Kebidanan	P	23	2.83	3.02	3.04	3.10
24	2019.132.001	Aina Salsabylla	D3 Farmasi	P	24	3.33	3.62	3.63	3.65
25	2019.132.002	Alfina Kusuma Sari	D3 Farmasi	P	23	2.17	2.17	2.17	2.17
26	2019.132.003	Alvinda Dewi Safitri	D3 Farmasi	P	24	2.75	2.79	2.91	2.82

Tabel 3. 3 (Lanjutan)

27	2019.132.004	Angelina Utama Mone	D3 Farmasi	P	23	3.33	3.43	3.39	3.40
28	2019.132.005	Apriani Paulina Tatck	D3 Farmasi	P	25	3.21	3.36	3.39	3.36
29	2019.132.006	Ayu Oktaviani	D3 Farmasi	P	23	3.92	3.77	3.70	3.62
30	2019.132.007	Daniatul Akbar	D3 Farmasi	L	22	2.50	2.68	2.73	2.72
31	2019.132.008	Diah Pu'an Maharani	D3 Farmasi	P	25	3.54	3.28	3.15	3.12
32	2019.132.009	Diah Nur Kholifah	D3 Farmasi	P	23	3.83	3.68	3.64	3.59
33	2019.132.010	Dian Nurcahyati	D3 Farmasi	P	23	3.67	3.66	3.63	3.62
34	2019.132.011	Dwi Anakasari	D3 Farmasi	P	24	3.38	3.51	3.60	3.62
35	2019.132.012	Dwi Wahyu Nengsih	D3 Farmasi	P	24	2.58	2.96	3.01	3.02
36	2019.132.013	Eiden Aerin Aktawolora	D3 Farmasi	P	22	2.50	0.00	0.00	0.00

Tabel 3. 4 Atribut Yang Digunakan

No	Atribut	Nilai
1	Program Studi	D3 Kebidanan, D3 Farmasi, D3 Rekam Medis dan Informasi Kesehatan, D3 Administrasi Rumah Sakit
2	Jenis Kelamin	Laki-Laki (L), Perempuan (P)
3	Usia	<=28 Tahun , >28 Tahun <=24 Tahun, >24 Tahun
4	Asal	Jawa, Luar Jawa
5	IPS1	<=3.39 >39 <=3.45 >3.45
6	IPS2	<=3.41 >3.41 <=3.43 > 3.43
7	IPS3	<=3.42 >3.42 <=3.45 >3.45
8	IPS3	<=3.43 >3.43

Jumlah data awal yang diperoleh dari pengumpulan data yaitu sebanyak 308 data, namun tidak semua data dapat digunakan dan tidak semua atribut digunakan karena harus melalui beberapa tahap pengolahan awal data

(*preparation data*). Untuk mendapatkan data yang berkualitas, beberapa teknik yang dilakukan adalah sebagai berikut (vecellis, 2009):

1. *Data validation,* untuk mengidentifikasi dan menghapus data yang ganjil (*outlier/noise*), data yang tidak konsisten, dan data yang tidak lengkap (*missing value*). *Missing data* terlihat pada tabel 4.1 dan hasilnya terlihat pada tabel 4.2.

2. *Data integration and Transformation,* untuk meningkatkan akurasi dan efisiensi algoritma. Data yang digunakan dalam penulisan ini bernilai kategorikal. Data ditransformasikan ke dalam *software RapidMiner*. Tabel kategorikal atribut terlihat pada tabel 3.4.

3. *Data size reduction and dicrtization,* untuk memperoleh data set dengan jumlah atribut dan record yang lebih sedikit tetapi bersifat informatif. Dalam penelitian ini atribut yang tidak relevan seperti nim, nama, jurusan, alamat, agama, indeks prestasi semester lima dan enam dihapuskan seperti terlihat pada tabel 3.2 dimana atribut yang digunakan menjadi tujuah atribut prediktor dan satu atribut *label*.

3.3. Metode Analisa data

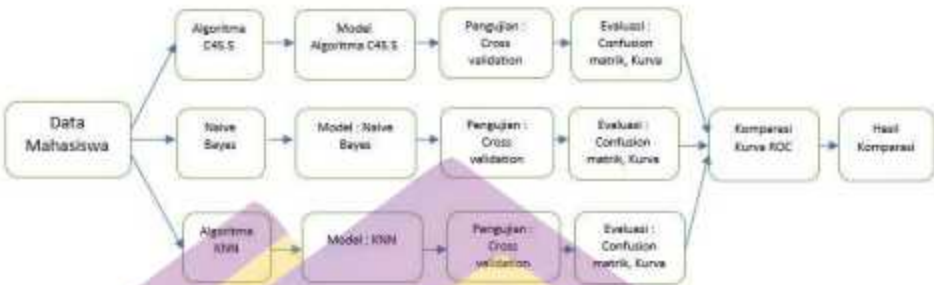
Pada penelitian ini dilakukan Analisa data tehadap data-data yang diperoleh pada tahap pengumpulan data. Sistem analisa data yang digunakan dalam penelitian ini yaitu mengunakan algoritma C4.5, Naïve Bayes dan K-Nearest Neighbors. Analisa data dilakukan dengan penentuan data set, dimana pengelolaan data set dilakukan dengan penentuan data set didapatkan dari bagian kemahasiswaan Poltekkes Permata Indonesia Yogyakarta.

Kemudian dilakukan penentuan atribut-atribut yang mempengaruhi kelulusan mahasiswa didapatkan berdasarkan hasil studi literature dan focus group discussion yang selanjutnya divalidasi oleh ahli/decision maker dalam hal ini adalah pengelola Program Studi dan bagian admisi atau penerimaan mahasiswa baru. Kemudian pembagian data set, dimana menyiapkan data traning dan data testing. Data traning ini akan digunakan dalam pembuatan C4.5, Naïve Bayes dan K-Nearest Neighbors.

3.4. Alur Penelitian

Dalam penelitian ini akan dilakukan analisis komparasi menggunakan tiga metode klasifikasi data mining. Metode yang diusulkan untuk pengolahan data mahasiswa adalah pengunaan algoritma C4.5, *Naïve Bayes* dan *K-Nearest Neighbors*. Data diolah sesuai dengan algoritmanya masing-masing, yakni data mahasiswa diolah menggunakan metode algoritma C4.5, *Naïve Bayes* dan *KNearest Neighbors*, setelah diolah dan menghasilkan model, maka terhadap model yang dihasilkan tersebut dilakukan pengujian menggunakankan *k-fold cross validation*, kemudian dilakukan evaluasi dan validasi hasil dengan *confusion matrix* dan kurva *ROC*. Tahap selanjutnya adalah memperbandingkan hasil akurasi dan AUC dari setiap model, sehingga diperoleh.

Model dari metode klasifikasi yang mana yang memperoleh nilai akurasi dan AUC tertinggi. Dalam tahapan ini akan dilakukan beberapa langkah-langkah metode yang diusulkan data yaitu seperti pada gambar 3.1 berikut:



Gambar 3. 1 . Alur Penelitian

Hasil pengujian dengan akurasi yang paling tinggi adalah metode yang akan digunakan untuk prediksi kelulusan mahasiswa ini. Berikut gambaran kateristik dari masing-masing metode:

a. Algoritma C4.5 yaitu salah satu algoritma dalam metode *decision tree* yang merubah data menjadi pohon keputusan menggunakan rumus perhitungan entropi.

b. *Naïve bayes* yaitu metode yang menghitung probabilitas antara kemunculan data yang satu dengan data yang lainnya.

c. *K-Nearest Neighbors* Algoritma ini menggunakan klasifikasi terhadap suatu objek berdasarkan data jarak tetangga *(neighbor)* terdekatnya.

BAB IV

HASIL PENELITIAN DAN PEMBAHASAN

4.1. Hasil

Pada penelitian ini akan digunakan 3 algoritma data mining yaitu C4.5 atau yang biasa disebut dengan Decision Tree, Naive Bayes dan KNN (K-Nearest Neighbor). Ketiga algoritma tersebut akan dievaluasi dengan membandingkan hasil dari kedua algoritma dengan nilai confusion matrix yaitu nilai accuracy dan AUC.

Data yang di gunakan dalam penelitian ini sebanyak 320 Data yang tidak ada nilainya atau bernilai tersebut dihilangkan sehingga dari 320 *record* diperoleh 305 *record*. Dari 305 data, jumlah data yang lulus "TEPAT WAKTU" yakni sebanyak 252 data dan jumlah data yang lulus "TERLAMBAT" sebanyak 53 data. Sampel data training yang digunakan seperti terlihat pada tabel 4.1, dan tabel pengkategorian setelah dilakukan teknik *integration* dan *tranformation* terlihat pada tabel 4.2.

Tabel 4. 1 Ilustrasi Missing Data Pada Data Training

Prodi	Jenis Kelamin	Umur	IPS1	IPS2	IPS3	IPS4	Status
D3 Farmasi	P	24	2.58	2.96	3.01	3.02	TEPAT WAKTU
D3 Farmasi	P	22	2.50	0.00	0.00	0.00	TERLAMBAT
D3 Farmasi	P	23	3.79	3.79	3.75	3.67	TEPAT WAKTU
D3 Farmasi	P	22	3.17	3.09	3.12	3.20	TEPAT WAKTU
D3 Farmasi	P	24	2.42	2.70	2.82	2.96	TEPAT WAKTU
D3 Farmasi	P	25	2.83	3.19	3.16	3.20	TEPAT WAKTU
D3 Farmasi	L	23	2.79	2.85	2.87	2.89	TEPAT WAKTU
D3 Farmasi	L	28	2.92	3.23	3.16	3.20	TEPAT WAKTU
D3 Farmasi	P	25	3.33	3.49	3.51	3.49	TEPAT WAKTU
D3 Farmasi	P	23	3.00	3.15	3.10	3.14	TEPAT WAKTU
D3 Farmasi	P	24	2.83	2.98	3.00	3.06	TEPAT WAKTU

Tabel 4. 1 (Lanjutan)

D3 Farmasi	P	23	3.00	3.21	3.24	3.32	TEPAT WAKTU
D3 Farmasi	P	23	3.67	3.72	3.75	3.69	TEPAT WAKTU
D3 Farmasi	P	22	3.33	3.45	3.45	3.44	TEPAT WAKTU
D3 Farmasi	P	23	3.13	3.34	3.40	3.42	TEPAT WAKTU
D3 Farmasi	P	22	2.50	2.98	3.15	3.25	TEPAT WAKTU
D3 Farmasi	L	23	2.75	2.74	2.85	2.87	TEPAT WAKTU
D3 Farmasi	P	25	3.17	3.30	3.40	3.38	TEPAT WAKTU
D3 Farmasi	P	26	3.92	3.83	3.78	3.72	TEPAT WAKTU
D3 Farmasi	P	23	3.04	3.28	3.34	3.40	TEPAT WAKTU
D3 Farmasi	P	24	2.92	3.02	3.06	3.06	TEPAT WAKTU
D3 Farmasi	P	22	2.67	2.70	2.76	2.78	TEPAT WAKTU
D3 Farmasi	P	21	2.83	2.83	2.94	3.01	TEPAT WAKTU
D3 Farmasi	P	22	3.08	3.19	3.19	3.21	TEPAT WAKTU
D3 Farmasi	P	22	3.17	2.98	3.15	3.18	TEPAT WAKTU
D3 Farmasi	P	24	3.79	3.72	3.75	3.74	TEPAT WAKTU
D3 Farmasi	P	22	3.17	2.91	2.42	2.42	TERLAMBAT
D3 Farmasi	P	22	2.75	3.04	3.12	3.15	TEPAT WAKTU
D3 Farmasi	P	23	3.88	3.77	3.75	3.74	TEPAT WAKTU
D3 Farmasi	L	25	2.67	2.98	2.99	2.98	TEPAT WAKTU
D3 Farmasi	P	22	3.04	3.00	3.06	3.11	TEPAT WAKTU
D3 Farmasi	L	23	2.63	3.17	3.19	3.19	TEPAT WAKTU
D3 Farmasi	P	23	3.29	3.36	3.45	3.48	TEPAT WAKTU
D3 Farmasi	P	24	3.17	3.23	3.33	3.36	TEPAT WAKTU
D3 Farmasi	P	23	3.04	3.09	3.19	3.26	TEPAT WAKTU
D3 Farmasi	P	25	2.92	2.98	3.07	3.06	TEPAT WAKTU
D3 Farmasi	P	24	3.13	3.19	3.24	3.25	TEPAT WAKTU
D3 Farmasi	L	23	0.00	0.00	0.00	0.00	TERLAMBAT
D3 Farmasi	P	23	3.00	3.17	3.33	3.34	TEPAT WAKTU
D3 Farmasi	P	24	2.58	2.96	3.01	3.02	TEPAT WAKTU
D3 Farmasi	P	22	2.50	0.00	0.00	0.00	TERLAMBAT
D3 Farmasi	P	23	3.79	3.79	3.75	3.67	TEPAT WAKTU

Tabel 4. 2 Data Traning

Program Studi	Jenis Kelamin	Umur	IPS1	IPS2	IPS3	IPS4
D3 Kebidanan	P	22	2.61	2.78	2.93	2.97
D3 Kebidanan	P	23	3.83	3.70	3.71	3.70
D3 Kebidanan	P	22	2.52	2.78	2.72	2.76

Tabel 4. 2 (Lanjutan)

D3 Kebidanan	P	24	2.35	2.72	2.72	2.83
D3 Kebidanan	P	22	2.48	2.65	2.72	2.85
D3 Kebidanan	P	23	3.13	3.22	3.18	3.16
D3 Kebidanan	P	23	2.83	3.02	3.04	3.10
D3 Farmasi	P	24	3.33	3.62	3.63	3.65
D3 Farmasi	P	23	2.17	2.17	2.17	2.17
D3 Farmasi	P	24	2.75	2.79	2.91	2.82
D3 Farmasi	P	23	3.33	3.43	3.39	3.40
D3 Farmasi	P	23	3.29	3.36	3.45	3.48
D3 Farmasi	P	24	3.17	3.23	3.33	3.36
D3 Farmasi	P	23	3.04	3.09	3.19	3.26
D3 Farmasi	P	25	2.92	2.98	3.07	3.06
D3 Farmasi	P	24	3.13	3.19	3.24	3.25
D3 Farmasi	P	23	3.00	3.17	3.33	3.34
D3 Farmasi	P	26	3.63	3.62	3.64	3.67
D3 Farmasi	P	26	3.71	3.70	3.72	3.71
D3 Farmasi	P	24	3.46	3.57	3.63	3.66
D3 Farmasi	P	27	3.46	3.62	3.63	3.59
D3 Farmasi	P	23	3.92	3.81	3.72	3.73
D3 Farmasi	P	24	3.71	3.53	3.39	3.42
D3 Farmasi	P	27	3.79	3.74	3.78	3.75
D3 Farmasi	P	24	3.63	3.40	3.37	3.41
D3 Farmasi	P	28	3.92	3.87	3.79	3.79
D3 Farmasi	P	26	4.00	3.83	3.78	3.78
D3 Farmasi	P	25	3.58	3.60	3.63	3.66
D3 Farmasi	L	27	3.46	3.55	3.57	3.58
D3 Farmasi	P	25	3.71	3.70	3.69	3.73
D3 Farmasi	P	26	3.71	3.70	3.67	3.69
D3 Farmasi	P	25	3.83	3.79	3.73	3.67
D3 Farmasi	P	25	3.50	3.49	3.52	3.53
D3 Farmasi	P	24	3.63	3.62	3.58	3.65
D3 Farmasi	P	27	3.63	3.66	3.69	3.73
D3 Farmasi	L	41	3.92	3.62	3.62	3.62
D3 Farmasi	P	25	3.46	3.57	3.63	3.66
D3 Farmasi	P	24	3.54	3.66	3.76	3.74
D3 Farmasi	P	47	3.50	3.32	3.15	3.18
D3 Farmasi	L	71	3.13	3.30	3.13	3.20
D3 Farmasi	P	37	3.17	3.32	3.21	3.22
D3 Farmasi	P	36	3.25	3.13	3.18	3.22

Tabel 4. 2 (Lanjutan)

D3 Farmasi	P	44	3.42	3.30	3.19	3.19
D3 Farmasi	L	31	3.13	3.11	3.03	3.08
D3 Rekam Medis Dan Informasi Kesehatan	P	22	3.55	3.57	3.71	3.76
D3 Rekam Medis Dan Informasi Kesehatan	L	23	3.55	3.48	3.65	3.69
D3 Rekam Medis Dan Informasi Kesehatan	P	22	2.91	3.04	3.14	3.16
D3 Rekam Medis Dan Informasi Kesehatan	L	27	3.18	2.83	2.83	2.80
D3 Rekam Medis Dan Informasi Kesehatan	L	29	3.09	3.35	3.39	3.34
D3 Rekam Medis Dan Informasi Kesehatan	P	23	3.73	3.70	3.67	3.66
D3 Rekam Medis Dan Informasi Kesehatan	L	23	3.36	3.57	3.58	3.56
D3 Rekam Medis Dan Informasi Kesehatan	P	27	3.55	3.48	3.51	3.44
D3 Rekam Medis Dan Informasi Kesehatan	P	23	3.91	3.87	3.86	3.82
D3 Rekam Medis Dan Informasi Kesehatan	P	23	3.64	3.70	3.71	3.71
D3 Rekam Medis Dan Informasi Kesehatan	P	24	3.82	3.78	3.86	3.89
D3 Rekam Medis Dan Informasi Kesehatan	P	23	3.09	3.39	3.54	3.56
D3 Rekam Medis Dan Informasi Kesehatan	P	24	4.00	3.91	3.91	3.87
D3 Rekam Medis Dan Informasi Kesehatan	P	23	3.64	3.70	3.77	3.82
D3 Rekam Medis Dan Informasi Kesehatan	L	22	3.64	3.70	3.77	3.82
D3 Rekam Medis Dan Informasi Kesehatan	L	22	3.73	3.74	3.80	3.82
D3 Rekam Medis Dan Informasi Kesehatan	P	23	3.64	3.70	3.80	3.84
D3 Rekam Medis Dan Informasi Kesehatan	P	23	3.55	3.61	3.65	3.64
D3 Rekam Medis Dan Informasi Kesehatan	P	25	3.73	3.78	3.77	3.76
D3 Rekam Medis Dan Informasi Kesehatan	P	24	3.82	3.78	3.80	3.84
D3 Rekam Medis Dan Informasi Kesehatan	P	23	3.55	3.43	3.54	3.56

Tabel 4. 2 (Lanjutan)

D3 Rekam Medis Dan Informasi Kesehatan	P	22	3.73	3.61	3.62	3.59
D3 Rekam Medis Dan Informasi Kesehatan	P	23	3.64	3.57	3.68	3.76
D3 Rekam Medis Dan Informasi Kesehatan	P	23	3.55	3.48	3.49	3.44
D3 Rekam Medis Dan Informasi Kesehatan	P	23	3.64	3.65	3.62	3.58
D3 Rekam Medis Dan Informasi Kesehatan	P	22	3.55	3.65	3.67	3.72
D3 Rekam Medis Dan Informasi Kesehatan	P	25	3.45	3.43	3.51	3.43
D3 Rekam Medis Dan Informasi Kesehatan	P	23	3.45	3.61	3.59	3.56
D3 Rekam Medis Dan Informasi Kesehatan	P	22	3.09	3.26	3.23	3.19
D3 Rekam Medis Dan Informasi Kesehatan	L	30	3.45	3.57	3.68	3.69
D3 Rekam Medis Dan Informasi Kesehatan	L	48	3.82	3.87	3.91	3.84
D3 Rekam Medis Dan Informasi Kesehatan	P	26	3.55	3.57	3.68	3.72
D3 Rekam Medis Dan Informasi Kesehatan	P	26	3.09	3.17	3.01	3.01
D3 Rekam Medis Dan Informasi Kesehatan	P	45	3.91	3.91	3.94	3.96
D3 Rekam Medis Dan Informasi Kesehatan	P	23	3.64	3.61	3.74	3.77
D3 Rekam Medis Dan Informasi Kesehatan	P	23	3.73	3.61	3.65	3.64
D3 Rekam Medis Dan Informasi Kesehatan	P	26	3.45	3.43	3.54	3.52
D3 Rekam Medis Dan Informasi Kesehatan	P	27	3.64	3.70	3.77	3.79
D3 Rekam Medis Dan Informasi Kesehatan	P	22	3.09	3.17	3.29	3.33
D3 Administrasi Rumah Sakit	P	27	3.67	3.79	3.85	3.86
D3 Administrasi Rumah Sakit	P	23	3.75	3.83	3.88	3.88
D3 Administrasi Rumah Sakit	P	24	3.00	3.31	3.42	3.29

Tabel 4. 2 (Lanjutan)

D3 Administrasi Rumah Sakit	P	22	3.42	3.58	3.67	3.69
D3 Administrasi Rumah Sakit	P	23	3.50	3.71	3.76	3.81
D3 Administrasi Rumah Sakit	P	22	3.33	3.58	3.67	3.71
D3 Administrasi Rumah Sakit	P	22	3.25	3.54	3.67	3.71

4.2. Model Proses Komparasi Decision Tree, Naïve Bayes, dan K-NN

Data training pada Tabel 4.2. adalah untuk menentukan apakan seseorang mahasiswa lulus tepat waktu atau terlambat. Berikut akan dibahas langkah-langkah perhitungan prediksi mahasiswa lulus tepat waktu atau tidak dengan menggunakan algoritma C4.5. Berikut langkah dalam pembuatan pohon keputusan, yaitu:

1. Menyiapkan data training, untuk data training yang digunakan ada pada tabel 4.2.
2. Hitung nilai entropy keseluruhan total kasus "TEPAT WAKTU" lulus dan "TERLAMBAT" lulus. Dari data training yang ada diketahui jumlah kasus yang lulus "TEPAT WAKTU" pada waktunya sebanyak 252 record, dan jumlah kasus yang lulus "TERLAMBAT" adalah sebanyak 53 record total kasus keseluruhan adalah 305 kasus. Sehingga didapat entropy keseluruhan pada gambar 4.1 :

$Entropy\ (S) = \sum_{i=1}^{n} - pi * Log2\ pi$ .....(4.1)

= -252/305 *log 252/305 + (-53/305 * log 53/305)

= 0.6661

H(S) = - (252 305 log2(252 305) + 53 305 log2 (53 305))

Jika kita substitusi nilai ke dalam ekspresi tersebut:

H(S) = - (252 305 log2(252 305) + 53 305 log2 (53 305))

H(S) ≈ - (252 305 x (-0.1072) + 53 305 x (-2.9367))

H(S) ≈ - (-0.0877 + 0.5122)

H(S) ≈ -0.5999

Jadi, nilai entropy keseluruhan (H(S)) sekitar 0.5999. Ini adalah ukuran ketidakpastian atau kejanggalan dalam dataset "TEPAT WAKTU" dan "TERLAMBAT"

3. Hitung nilai entropi dan nilai gain masing-masing atribut. Nilai gain tertinggi adalah atribut yang menjadi root dari pohon keputusan yang akan dibuat.

Misalkan menghitung entropi bagi atribut Program Studi Kebidanan pada Gambar 4.2. nilai Entropi Pada Farmasi pada gambar 4.3., Entropi Pada Rekam Medis dan Informasi Kesehatan pada gambar 4.4 dan nilai Entropi pada Administrasi Rumah sakit pada gambar 4.5

EKebidanan [12,11] = (-12/23 log2 12/23)+(-11/23 log2 11/23)

H(S)= -(log2(12/23)+log2(11/23)) Mari hitung nilai ini:

H(S) ≈ - (12/23 × (-0.355)+11/23 × (-0.389))

H(S) ≈ - (-0.185 + (-0.174)) H(S) ≈ -(-0.359)

H(S) ≈ 0.359

EFarmasi[129,34]=(-129/163 log(2) 129/163) + (-34/163 log(2) 34/163)

H(S) = - (129/163 log2 (129/163) + 34/163 log2 (34/163))

Mari hitung nilai ini:

H(S) ≈ - (129/163 × (-0.678) + 34/163 × (-1.880))

H(S) ≈ - (-0.535 + (-0.396)) H(S) ≈ -(-0.931)

H(S) ≈ 0.931

ERekam Medis & Infokes[99,6]=(-99/105 log(2) 99/105) + (-6/105 log(2) 6/105)

H(S) = - (99/105 log2(99/105) + 6/105 log2(6/105)) Mari hitung nilai ini:

H(S) ≈ - (99/105 × (-0.151) + 6/105 × (-2.807))

H(S) ≈ - (-0.142 + (-0.162)) H(S) ≈ -(-0.304)

H(S) ≈ 0.304

Jadi, nilai entropy (H(S)) untuk kasus "ERekam Medis & Infokes" sekitar 0.304.

EAdministrasi Rumah Sakit[12,2]=(-12/14 log(2) 12/14) + (-2/14 log(2) 2/14)

H(S) = –(12/14 log2(12/14) + 2/14 log2(2/14))

Mari hitung nilai ini:

H(S) ≈ –(12/14 × (–0.139) + 2/14 × (–2.807))

H(S) ≈ –(–0.118 + (–0.401)) H(S) ≈ –(–0.519)

H(S) ≈ 0.519

Jadi, nilai entropy (H(S)) untuk kasus "EAdministrasi Rumah Sakit" sekitar 0.519.

2. Kemudian hitung gain Program Studi seperti berikut:

$$Gain\ (S,A)Entropy(S) - \sum_{i=0}^{n} \square \frac{|S|}{|S|} * Entropy(S) \quad \text{.....(4.2)}$$

Gain(S,A)= 0,666-((23/305 * 0,998)+(163/305 * 0,738)+(105/305 * 0,316)+(14/305 * 0,592) = 0.0602

Gain Prodi=0.0602

Berikut table 4.1 hasil nilai entropi dan *gain* dari masing-masing atribut:

Tabel 4. 1 Nilai entropy dan gain untuk menentukan akar

Simpul	Kasus	Tepat Waktu	Terlambat	Entropi	Gain
Jumlah Kasus	305	252	53	0.666	
Program Studi					0.0602
D3 Kebidanan	23	12	11	0.998	
D3 Farmasi	163	129	34	0.738	
D3 Rekam Medis & Infokes	105	99	6	0.316	
D3 Administrasi RS	14	12	2	0.592	
Jenis Kelamin					0.0107
Laki-Laki	40	28	12	0.881	
Perempuan	265	224	41	0.621	
Asal					0.0099
Jawa	199	171	28	0.586	
Luar Jawa	106	81	25	0.788	
Umur					

Tabel 4. 1 (Lanjutan)

<=28 tahun	223	182	41	0.68844135	0.001430518
>28 Tahun	82	70	12	0.600608575	
<=24 Tahun	155	122	33	0.746995624	0.008025851
>24 Tahun	150	130	20	0.566509507	
IPS1					
<=3.39	135	99	36	0.836640742	0.034534036
>3.39	170	153	17	0.468995594	
<=3.45	142	103	39	0.848055283	0.045577611
>3.45	163	149	14	0.422598839	
IPS2					
<=3.41	147	100	47	0.904085304	0.10985407
>3.41	158	152	6	0.232927855	
<=3.43	152	105	47	0.892260044	0.101856968
>3.43	153	147	6	0.238684511	
IPS3					
<=3.42	148	101	47	0.901716637	0.10822296
>3.42	157	151	6	0.234054566	
<=3.45	158	111	47	0.878174184	0.092759974
>3.45	147	141	6	0.246022578	
IPS4					
<=3.43	151	104	47	0.894620589	0.103425144
>3.43	154	148	6	0.237508144	

Dari tabel 4.1 dapat dilihat nilai *gain* tertinggi ada pada atribut IPS2 yakni 0.10985407 sehingga dapat atribut IPS2 adalah akar dari pohon keputusan. Kemudian dilakukan kembali perhitungan nilai entropi dan *gain* untuk menentukan simpul 1.1, nilai yang dihitung berdasarkan atribut IPS2. Perhitungan nilai *entropy* dan *gain* untuk atribut=bahasa dan seni disajikan dalam tabel 4.2.

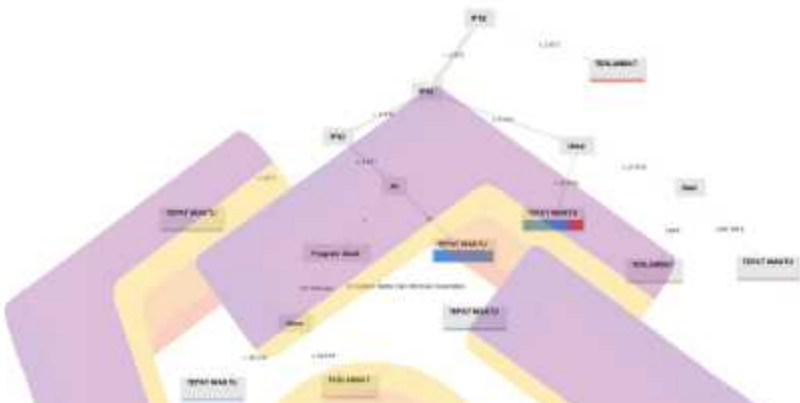
Tabel 4. 2 Tabel nilai entropy dan gain untuk menentukan simpul 1.1

Simpul	Kasus	Tepat Waktu	Terlambat	Entropi	Gain
IPS2	147	100	47	0.904085304	
Program Studi					0.030065865
D3 Kebidanan	18	8	10	0.99107606	
D3 Farmasi	29	23	6	0.735508582	
D3 Rekam Medis & Infokes	97	67	30	0.89233868 1	
D3 Administrasi RS	3	2	1	0.918295834	
Jenis Kelamin					0.030065865
Laki-Laki	26	16	10	0.961236605	
Perempuan	121	84	37	0.888248047	
Asal					0.016764959
Jawa	85	63	22	0.824965868	
Luar Jawa	62	37	25	0.972806322	
Umur					
<=22 tahun	9	3	6	0.918295834	0.023882177
>22 Tahun	138	97	41	0.87771882	
<=28 Tahun	89	53	36	0.973519002	0.038189623
>28 Tahun	58	47	11	0.70074955	

Tabel 4.2 (Lanjutan)

IPS1					
<=3.39	114	78	36	0.899743759	0.000176794
>3.39	33	22	11	0.918295834	
<=3.45	119	80	39	0.912591426	0.000916898
>3.45	28	20	8	0.863120569	
IPS2					
<=3.43	147	100	47	0.904085304	0
>3.43	0	0	0	0	
<=3.41	147	100	47	0.904085304	0
>3.41	0	0	0	0	
IPS3					
<=3.42	135	89	46	0.92552578	0.020331682
>3.42	12	11	1	0.41381685	
<=3.45	141	95	46	0.911039772	0.003699301
>3.45	6	5	1	0.650022422	
IPS4					
<=3.43	136	90	46	0.923119984	0.017154796
>3.43	11	10	1	0.439496987	

Untuk semua atribut dan nilainya lakukan cara yang sama, sehingga diperoleh pohon keputusan seperti gambar 4.1.



Gambar 4. 1 Pohon keputusan data kelulusan mahasiswa

Gambar 4.1 diatas menunjukkan hasil deskripsi secara lengkap dari pohon keputusan yang telah terbentuk dengan menggunakan Algoritma C4.5. Dari hasil deskripsi juga menunjukkan bahwa Algoritma C4.5 baik digunakan dalam proses penggalian data untuk menarik beberapa kesimpulan yang divisualisasikan ke dalam pohon keputusan. Berikut merupakan rule yang diperoleh pohon keputusan yang terbentuk.

1. Langkah pertama:
   Jika IPS2 lebih besar dari 2.872, kita pergi ke langkah berikutnya.
   Jika IPS2 kurang dari atau sama dengan 2.872, prediksi TERLAMBAT.
2. Langkah kedua:
   Jika IPS2 lebih besar dari 3.410, kita pergi ke cabang berikutnya.
   Jika IPS2 kurang dari atau sama dengan 3.410, kita pergi ke cabang terakhir.
3. Cabang pertama dari Langkah kedua:

Jika IPS2 lebih besar dari 3.871, prediksi TEPAT WAKTU dengan distribusi 4 ke TEPAT WAKTU dan 1 ke TERLAMBAT.

Jika IPS2 kurang dari atau sama dengan 3.871, kita pergi ke cabang berikutnya.

4. Cabang kedua dari Langkah kedua:

   Jika Jenis Kelamin (JK) adalah Laki-Laki (L):

   Jika Program Studi = D3 Farmasi dan Umur lebih besar dari 32.019, prediksi TEPAT WAKTU dengan distribusi 3 ke TEPAT WAKTU dan 0 ke TERLAMBAT.

   Jika Program Studi = D3 Farmasi dan Umur kurang dari atau sama dengan 32.019, prediksi TERLAMBAT dengan distribusi 0 ke TEPAT WAKTU dan 2 ke TERLAMBAT.

   Jika Program Studi = D3 Rekam Medis Dan Informasi Kesehatan, prediksi TEPAT WAKTU dengan distribusi 9 ke TEPAT WAKTU dan 0 ke TERLAMBAT.

   Jika JK = P (Perempuan), prediksi TEPAT WAKTU dengan distribusi 136 ke TEPAT WAKTU dan 3 ke TERLAMBAT.

5. Cabang ketiga dari Langkah kedua:

   Jika Umur lebih besar dari 21.916, prediksi TEPAT WAKTU dengan distribusi 98 ke TEPAT WAKTU dan 30 ke TERLAMBAT.

   Jika Umur kurang dari atau sama dengan 21.916, kita pergi ke cabang terakhir.

6. Cabang terakhir dari Langkah kedua:

   Jika Asal = Jawa, prediksi TERLAMBAT dengan distribusi 0 ke TEPAT WAKTU dan 2 ke TERLAMBAT.

   Jika Asal = Luar Jawa, prediksi TEPAT WAKTU dengan distribusi 2 ke TEPAT WAKTU dan 2 ke TERLAMBAT.

7. Langkah terakhir:

   Jika IPS2 kurang dari atau sama dengan 2.872, prediksi TERLAMBAT dengan distribusi 0 ke TEPAT WAKTU dan 13 ke TERLAMBAT.

4.1.2. *Naïve Bayes*

Data *training* yang digunakan untuk metode *naïve bayes* menggunakan data pada table 4.2. Dengan mencari *prior probability* untuk nilai yang Tepat waktu dan terlambat untuk semua jumlah data. Jika diketahui dalam data *training*, jumlah data 305, mahasiswa yang lulus dengang tepat waktu sebanyak 252 *record* dan mahasiswa yang lulus terlambat sebanyak 53 record. Berikut hasil perhitungan *prior probability* dengan menggunakan rumus (2.5) dan (2.6) :

P(Tepat Waktu,n) = 252/305 = 0.826

P(Terlambat,n) = 52/305= 0.2

Setelah itu mencari masing-masing setiap *class* atribut. Berikut hasil perhitungan *prior probability* untuk D3 Kebidanan dalam katagori Tepat Waktu :

$$P(x|y) = \frac{P(y|x)\,P(x)}{P(y)} \qquad (4.4)$$

P(Program_Studi=D3 Kebidanan|tepat) =12/671=0,192250373.

Berikut hasil perhitungan *priori probability* untuk masing-masing atribut, terdapat pada tabel 4.3:

Tabel 4.3. Perhitungan nilai probabilitas prior

Simpul	Tepat Waktu	Terlambat	P(X|Ci)	
			Tepat Waktu	Terlambat
Jumlah Kasus	252	53	0.826	0.2
Program Studi				
D3 Kebidanan	12	11	0.0476	0.2075
D3 Farmasi	129	34	0.5119	0.641
D3 Rekam Medis & Infokes	99	6	0.3928	0.113
D3 Administrasi RS	12	2	0.0476	0.0377
Jenis Kelamin				
Laki-Laki	28	12	0.111	0.2264
Perempuan	224	41	0.888	0.773
Asal				
Jawa	171	28	0.678571429	0.528301887
Luar Jawa	81	25	0.321428571	0.471698113
Umur				
<=22 tahun	12	6	0.047619048	0.113207547
>22 Tahun	240	47	0.952380952	0.886792453
<=28 Tahun	182	41	0.722222222	0.773584906
>28 Tahun	70	12	0.277777778	0.226415094
IPS1				
<=3.45	142	45	0.563492063	0.849056604
>3.45	110	8	0.436507937	0.150943396
<=3.39	99	37	0.392857143	0.698113208
>3.39	153	16	0.607142857	0.301886792
IPS2				
<=3.41	8	16	0.031746032	0.301886792
>3.41	244	37	0.968253968	0.698113208
<=3.43	95	45	0.376984127	0.849056604
>3.43	157	8	0.623015873	0.150943396
IPS3				
<=3.42	101	47	0.400793651	0.886792453
>3.42	151	6	0.599206349	0.113207547
<=3.45	111	47	0.44047619	0.886792453
>3.45	141	6	0.55952381	0.113207547
IPS4				
<=3.43	104	47	0.412698413	0.886792453
>3.43	148	6	0.587301587	0.113207547

Untuk menetukan kelas dari kasus baru maka dilakukan perhitungan *probabilitas posterior* berdasarkan probabilitas *prior* yang telah dihitung sebelumnya dan disajikan pada Tabel 4.2. perhitungan probabilitas *posterior* untuk menentukan data *testing* termasuk klasifikasi yang mana, sebagai contoh diambil kasus seperti Tabel 4.3, dimana data X tersebut adalah data yang akan diprediksi kelulusannya.

Tabel 4.4. Atribut X yang akan diprediksi

Data X untuk kasus terbaru	
Atribut	Nilai
Program Studi	D3 Farmasi
Jenis kelamin	Perempuan
Umur	23
Asal	Jawa
IPS1	2.87
IPS2	3,34
IPS3	3,8
IPS4	3.43

Berdasarkan nilai probabilitas prior masing-masing atribut yang telah dihitung pada Gambar 4.7 maka dapat dilihat rule yang diperoleh untuk atribut diatas yaitu seperti dibawah ini:

1. Hitung probabilitas TEPAT WAKTU untuk setiap atribut

   Dengan menggunakan rumus : $P(x|y) = P(y|x)\ P(x)$ (4.5)

P(tepat waktu)P(D3 Farmasi |tepat waktu)P(perempuan|tepat waktu)

2. Hitung probabilitas TERLAMBAT untuk setiap atribut
   Dengan menggunakan rumus: P(x|y) = P(y|x) P(x) (4.6)

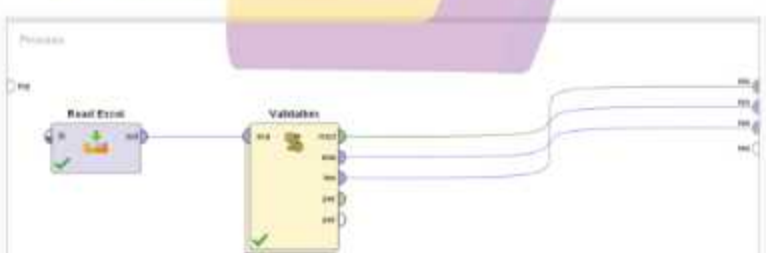
3. Bandingkan hasil dari probabilitas TEPAT Waktu dan TERLAMBAT

Dikarenakan 4.7433E+183 > 2.12225E-07 maka dapat untuk data *testing* yang ada pada tabel 4.2 termasuk kelas TERLAMBAT.

Rule1: Jika probabilitas tepat lebih besar dari probabilitas terlambat maka hasil adalah TEPAT WAKTU

Rule2: Jika probabilitas terlambat lebih besar dari probabilitas tepat maka hasil adalah TERLAMBAT.

Setelah diolah maka dilakukan teknik pengujian dengan metode *k-fold cross validation* pada *tools RapidMiner*, pegolahan pengujian untuk metode algoritma *C4.5, Naive Bayes dan K-Nearest Neighbor* terlihat seperti gambar 4.2 dibawah ini:



Gambar 4. 2 . Model Proses Desain Import Data

1. Read Excel : Operator ini dapat digunakan untuk memuat data dari spreadsheet Microsoft Excel.
2. Cross Validation : Operator yang bersarang. Ini memiliki dua subproses: subproses pelatihan dan subproses pengujian. Subproses pelatihan digunakan untuk melatih model. Model yang terlatih kemudian diterapkan dalam subproses pengujian. Kinerja model diukur selama fase Pengujian

Gambar 4.3 Model Model Validasi C4 .5

Gambar 4.4 Model Model Validasi Naïve Bayes

Gambar 4.5 Model Model Validasi KNN

3. Model Validasi: Metode klasifikasi yang digunakan dalam penelitian ini yaitu Decision Tree dapat dilihat pada gambar 4.2, Naive Bayes Dapat dilihat pada gambar 4.3, dan K-Nearest Neighbor dapat dilihat pada gambar 4.4.
4. Apply Model : Operator yang digunakan untuk penghubung metode Decision Tree, Naive Bayes dan K-Nearest Neighbor ke performance. Dapat dilihat pada gambar 4.3,4.4 dan 4.5.
5. Performance : Operator yang digunakan untuk mengukur performance akurasi dari model.

4.2. Evaluasi dan Validasi Metode

Metode klasifikasi bisa dievaluasi berdasarkan kriteria seperti tingkat akurasi, kecepatan, kehandalan, skabilitas dan interpretabilitas (Vecellis, 2009). Setelah data diolah maka dapat diuji tingkat akurasinya untuk melihat kinerja dari masing-masing metode.

Penelitian ini bertujuan untuk melihat akurasi analis kelulusan mahasiswa pada suatu universitas, menilai apakah dengan kriteria yang dimiliki mahasiswa dapat lulus tepat waktu atau tidak. Kemudian

melakukan perbandingan ketiga metode yakni algoritma C4.5, *naïve bayes* dan K-*Nearest Neighbor* kemudian menganalisa akurasi dengan membandingkan ketiga metode tersebut.

4.2.1. Hasil Evaluasi Model Decision Tree menggunakan Cross Validation dan Confusion Matrix


accuracy: 81.64% +/- 1.01% (micro average: 81.64%)

	true TEPAT WAKTU	true TERLAMBAT
pred. TEPAT WAKTU	232	36
pred. TERLAMBAT	20	17

Gambar 4.6 Confusion Matrix 1-fold cross validation

Dari proses evaluasi model Decision Tree pada gambar 4.6 dan proses validasi terbentuk hasil matrix akurasi sebesar 81.64 % pada pengujian ke 1.

Dibawah ini merupakan perhitungan akurasi menggunakan Confusion Matrix dari gambar 4.8 diatas.

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Akurasi = \frac{232 + 17}{232 + 17 + 20 + 36}$$

$$Akurasi = \frac{249}{305} = 0.816$$

$$Presisi = \frac{TP}{TP + FP}$$

$$Presisi = \frac{232}{232 + 20}$$

$$Presisi = \frac{232}{252} = 0.866$$

$$Recall = \frac{TP}{TP + FN}$$

$$Recall = \frac{232}{232 + 36}$$

$$Recall = \frac{232}{268} = 0.866$$

$$F1\ Score = \frac{2xPresisixRecall}{Presisi + Recall}$$

$$F1Score = \frac{2x0.921x0.866}{0.921 + 0.866}$$

$$F1\ Score = 0.892$$

accuracy: 81.64% +/- 1.01% (micro average: 81.64%)

	true TEPAT WAKTU	true TERLAMBAT
pred. TEPAT WAKTU	232	36
pred. TERLAMBAT	20	17

Gambar 4.7 Confusion Matrix 2-fold cross validation

Akurasi:

Akurasi = (TP+TN)/(TP+TN+FP+FN)

Akurasi = (243+14)/(243+14+9+39) Akurasi = 257/305 ≈ 0.842

Presisi (untuk "TEPAT WAKTU"):

Presisi = TP/(TP+FP) Presisi = 243/(243+9) Presisi = 243/252 ≈ 0.964

Recall (untuk "TEPAT WAKTU"):

Recall = TP/(TP+FN) Recall = 243/(243+39) Recall = 243/282 ≈ 0.862

F1 Score (untuk "TEPAT WAKTU"):

F1 Score = (2x Presisi x Recall)/(Presisi + Recall) F1 Score = (2×0.964×0.862)/(0.964+0.862) F1 Score ≈ 0.910

accuracy: 85.25% +/- 2.68% (micro average: 85.25%)

	true TEPAT WAKTU	true TERLAMBAT
pred. TEPAT WAKTU	242	35
pred. TERLAMBAT	10	18

Gambar 4.8 Confusion Matrix 3-fold cross validation

Berdasarkan gambar 4.8 memiliki presisi sekitar 0.9603, recall sekitar 0.8732, skor F1 sekitar 0.9149, dan akurasi sekitar 85.25%. Evaluasi ini memberikan gambaran tentang kinerja model dalam memprediksi kelas "TEPAT WAKTU" dan "TERLAMBAT".

1. Presisi (Precision):

Presisi = True Positive / (True Positive + False Positive) Presisi = 242 / (242 + 10) ≈ 0.9603

2. Recall (Sensitivitas):

Recall = True Positive / (True Positive + False Negative) Recall = 242 / (242 + 35) ≈ 0.8732

3. Skor F1:

F1 = 2 × (Presisi × Recall) / (Presisi + Recall) F1 = 2 × (0.9603 × 0.8732) / (0.9603 + 0.8732) ≈ 0.9149

4. Akurasi:

Akurasi = (True Positive + True Negative) / Total Akurasi = (242 + 18) / (242 + 35 + 10 + 18) = 260 / 305 ≈ 0.8525

accuracy: 83.93% +/- 3.48% (micro average: 83.93%)

	true TEPAT WAKTU	true TERLAMBAT
pred. TEPAT WAKTU	236	33
pred. TERLAMBAT	16	20

Gambar 4.9 Confusion Matrix 4-fold cross validation

Presisi (Precision):

Presisi = True Positive / (True Positive + False Positive)

Presisi = 236 / (236+16) ≈ 0.9365

Recall (Sensitivitas):

Recall = True Positive / (True Positive + False Negative)

Recall = 236 / (236+33) ≈ 0.8774

Skor F1:

F1 = 2 * (Presisi * Recall) / (Presisi + Recall)

F1 = 2 * (0.9365 * 0.8774) / (0.9365 + 0.8774) ≈ 0.9061

Akurasi:

Akurasi = (True Positive + True Negative) / Total

Akurasi = (236 + 20) / (236 + 33 + 16 + 20) ≈ 256 / 305 ≈ 0.8393

Berdasarkan Gambar 4.9. memiliki presisi sekitar 0.9365, recall sekitar 0.8774, skor F1 sekitar 0.9061, dan akurasi sekitar 83.93%. Evaluasi ini memberikan gambaran tentang kinerja model dalam memprediksi kelas "TEPAT WAKTU" dan "TERLAMBAT".

accuracy: 83.93% +/- 4.40% (micro average: 83.93%)

	true TEPAT WAKTU	true TERLAMBAT
pred. TEPAT WAKTU	237	34
pred. TERLAMBAT	15	19

Gambar 4.10 Confusion Matrix 5-fold cross validation

Presisi (Precision):

Presisi = True Positive / (True Positive + False Positive) Presisi = 237 / (237+15) ≈ 0.9405

Recall (Sensitivitas):

Recall = True Positive / (True Positive + False Negative) Recall = 237 / (237+34) ≈ 0.8744

Skor F1:

F1 = 2 * (Presisi x Recall) / (Presisi + Recall) F1 = 2 * (0.9405 x 0.8744) / (0.9405 + 0.8744) ≈ 0.9064

Akurasi:

Akurasi = (True Positive + True Negative) / Total Akurasi = (237+19) / (237+34+15+19) ≈ 256 / 305 ≈ 0.8393

Berdasarkan Gambar 4.10. table ini memiliki presisi sekitar 0.9405, recall sekitar 0.8744, skor F1 sekitar 0.9064, dan akurasi sekitar 83.93%. Evaluasi ini memberikan gambaran tentang kinerja model dalam memprediksi kelas "TEPAT WAKTU" dan "TERLAMBAT".

4.2.2. *Hasil Evaluasi Model Naïve Bayes menggunakan Cross Validation dan Confusion Matrix*

Gambar 4.11 Confusion Matrix 1-fold cross validation (NBC)

Akurasi:

Akurasi = (TP+TN) / (TP+TN+FP+FN) Akurasi = (215+28) / (215+28+37+25) Akurasi = 243 / 305 ≈ 0.797

Presisi (untuk "TEPAT WAKTU"):

Presisi = TP / (TP+FP) Presisi = 215 / (215+37) Presisi = 215 / 252 ≈ 0.854

Recall (untuk "TEPAT WAKTU"):

Recall = TP / (TP+FN) Recall = 215 / (215+25) Recall = 215 / 240 ≈ 0.896

F1 Score (untuk "TEPAT WAKTU"):

F1 Score = 2 x Presisi x Recall / (Presisi + Recall) F1 Score = (2 x 0.854 x 0.896) / (0.854 + 0.896) F1 Score ≈ 0.874

Pada Gambar 4.10 Confusion Matrik model memiliki akurasi sekitar 79.7%, presisi sekitar 85.4%, recall sekitar 89.6%, dan F1 score sekitar 87.4% untuk kelas "TEPAT WAKTU," serta akurasi sekitar 79.7%, presisi sekitar 52.8%, recall sekitar 43.1%, dan F1 score sekitar 47.5% untuk kelas "TERLAMBAT."

	true TEPAT WAKTU	true TERLAMBAT
pred. TEPAT WAKTU	215	25
pred. TERLAMBAT	37	28

Gambar 4.12 Confusion Matrix 2-fold cross validation (NBC)

Akurasi = (True Positive + True Negative) / Total Akurasi = (215+28) / (215+25+37+28) ≈ 243 / 305 ≈ 0.796

Presisi (Precision): Presisi = True Positive / (True Positive + False Positive) Presisi = 215 / (215+37) ≈ 0.8532

Recall (Sensitivitas): Recall = True Positive / (True Positive + False Negative) Recall = 215 / (215+25) = 0.8958

Skor F1: F1 = 2 * (Presisi * Recall) / (Presisi + Recall) F1 = 2 * (0.8532 * 0.8958) / (0.8532 + 0.8958) ≈ 0.8741

Akurasi sekitar 79.67%, presisi sekitar 85.32%, recall sekitar 89.58%, dan F1 score sekitar 87.4%

	true TEPAT WAKTU	true TERLAMBAT
pred. TEPAT WAKTU	218	23
pred. TERLAMBAT	34	30

Gambar 4.13 Confusion Matrix 3-fold cross validation (NBC)

Presisi (Precision): Presisi = True Positive / (True Positive + False Positive)

Presisi = 218 / (218+34) ≈ 0.8651

Recall (Sensitivitas): Recall = True Positive / (True Positive + False Negative) Recall = 218 / (218+23) ≈ 0.9041

Skor F1: F1 = 2 x (Presisix Recall) / (Presisi + Recall) F1 = 2 x (0.8651×0.9041) / (0.8651+0.9041) ≈ 0.8841

Akurasi: Akurasi = (True Positive + True Negative) / Total Akurasi = (218+30) / (218+23+34+30) ≈ 248 / 305 ≈ 0.8131

Berdasarkan Gambar 4.13 Confusion Matrix 3-fold cross validation (NBC) memiliki presisi sekitar 0.8651, recall sekitar 0.9041, skor F1 sekitar 0.8841, dan akurasi sekitar 81.31%.

Gambar 4.14 Confusion Matrix 4-fold cross validation (NBC)

Presisi = (True Positive) / (True Positive + False Positive) Presisi = 212 / (212 + 40) ≈ 0.8413

Recall = (True Positive) / (True Positive + False Negative) Recall = 212 / (212 + 23) ≈ 0.9026

Skor F1:

F1 = 2 * (Presisi * Recall) / (Presisi + Recall) F1 = 2 * (0.8413 * 0.9026) / (0.8413 + 0.9026) ≈ 0.8711

Akurasi = (True Positive + True Negative) / Total Akurasi = (212 + 30) / (212 + 23 + 40 + 30) = 242 / 305 ≈ 0.7934

Berdasarkan Gambar 4.14. Confusion Matrix 4-fold cross validation (NBC) memiliki presisi sekitar 0.8413, recall sekitar 0.9026, skor F1 sekitar 0.8711, dan akurasi sekitar 79.34%. Evaluasi ini memberikan gambaran tentang kinerja model dalam memprediksi kelas "TEPAT WAKTU" dan "TERLAMBAT"

Gambar 4.15 Confusion Matrix 5-fold cross validation (NBC)

Pada gambar 4.15 Confusion Matrix 5-fold cross validation (NBC) memiliki presisi sekitar 0.9405, recall sekitar 0.8744, skor F1 sekitar 0.9064, dan akurasi sekitar 83.93%. Evaluasi ini memberikan gambaran tentang kinerja model dalam memprediksi kelas "TEPAT WAKTU" dan "TERLAMBAT".

Presisi = True Positive / (True Positive + False Positive) Presisi = 237 / (237 + 15) ≈ 0.9405

Recall = True Positive / (True Positive + False Negative) Recall = 237 / (237 + 34) ≈ 0.8744

Skor F1:

F1 = 2 × (Presisi × Recall) / (Presisi + Recall) F1 = 2 × (0.9405 × 0.8744) / (0.9405 + 0.8744) ≈ 0.9064

Akurasi = (True Positive + True Negative) / Total Akurasi = (237 + 19) / (237 + 34 + 15 + 19) ≈ 256 / 305 ≈ 0.8393

4.2.3. *Hasil Evaluasi Model K-Nearest Neighbor menggunakan Cross Validation dan Confusion Matrix*

Gambar 4.16 Confusion Matrix 1-fold cross validation K-Nearest Neighbor

Proses evaluasi model K-Nearest Neighbor dan proses validasi terbentuk hasil matrix Accuracy sebesar 84.26% pada pengujian ke 4. Dibawah ini merupakan perhitungan akurasi menggunakan Confusion Matrix dari gambar 4.16 diatas.

Akurasi = (TP + TN) / (TP + TN + FP + FN) Akurasi = (243 + 14) / (243 + 14 + 9 + 39) Akurasi = 257/305 ≈ 0.842

Presisi (untuk "TEPAT WAKTU"):

Presisi = TP / (TP + FP) Presisi = 243 / (243 + 9) Presisi = 243 / 252 ≈ 0.964

Recall (untuk "TEPAT WAKTU"):

Recall = TP / (TP + FN) Recall = 243 / (243 + 39) Recall = 243 / 282 ≈ 0.862

F1 Score (untuk "TEPAT WAKTU"):

F1 Score = (2 x Presisi x Recall) / (Presisi + Recall) F1 Score = (2 x 0.964 x 0.862) / (0.964 + 0.862) F1 Score ≈ 0.910

Gambar 4.17 Confusion Matrix 2-fold cross validation K-Nearest Neighbor

Presisi = True Positive / (True Positive + False Positive) Presisi = 243 / (243 + 9) ≈ 0.9643

Recall = True Positive / (True Positive + False Negative) Recall = 243 / (243 + 39) ≈ 0.8617

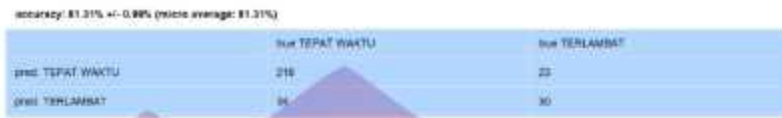
Skor F1:

F1 = 2 * (Presisi x Recall) / (Presisi + Recall) F1 = 2 * (0.9643 x 0.8617) / (0.9643 + 0.8617) ≈ 0.9101

Akurasi = (True Positive + True Negative) / Total Akurasi = (243 + 14) / (243 + 39 + 9 + 14) ≈ 257/305 ≈ 0.843

Berdasarkan Gambar 4.17 model diatas memiliki presisi sekitar 0.9643, recall sekitar 0.8617, skor F1 sekitar 0.9101, dan akurasi sekitar

84.3%. Evaluasi ini memberikan gambaran tentang kinerja model dalam memprediksi kelas "TEPAT WAKTU" dan "TERLAMBAT".



accuracy: 81.31% +/- 0.89% (micro average: 81.31%)

	true TEPAT WAKTU	true TERLAMBAT
pred. TEPAT WAKTU	218	23
pred. TERLAMBAT	34	30

Gambar 4.18 Confusion Matrix 3-fold cross validation K-Nearest Neighbor

Presisi = True Positive / (True Positive + False Positive) Presisi = 218 / (218 + 34) ≈ 0.8651

Recall = True Positive / (True Positive + False Negative) Recall = 218 / (218 + 23) ≈ 0.9041

F1 = 2 * (Presisi * Recall) / (Presisi + Recall) F1 = 2 * (0.8651 * 0.9041) / (0.8651 + 0.9041) ≈ 0.8841

Akurasi = (True Positive + True Negative) / Total Akurasi = (218 + 30) / (218 + 23 + 34 + 30) ≈ 248 / 305 ≈ 0.8131

Berdasarkan Gambar 4.18 diatas memiliki presisi sekitar 0.8651, recall sekitar 0.9041, skor F1 sekitar 0.8841, dan akurasi sekitar 81.31%. Evaluasi ini memberikan gambaran tentang kinerja model dalam memprediksi kelas "TEPAT WAKTU" dan "TERLAMBAT".

accuracy: 83.28% +/- 2.71% (micro average: 83.28%)

	true TEPAT WAKTU	true TERLAMBAT
pred. TEPAT WAKTU	243	42
pred. TERLAMBAT	9	11

Gambar 4.19 Confusion Matrix 4-fold cross validation K-Nearest Neighbor

Presisi (Precision): Presisi = True Positive / (True Positive + False Positive)

Presisi = 243 / (243+9) ≈ 0.9643

Recall (Sensitivitas): Recall = True Positive / (True Positive + False Negative)

Recall = 243 / (243+42) ≈ 0.8521

Skor F1: F1 = 2 * (Presisi * Recall) / (Presisi + Recall) F1 = 2 * (0.9643 * 0.8521) / (0.9643 + 0.8521) ≈ 0.9042

Akurasi: Akurasi = (True Positive + True Negative) / Total Akurasi = (243+11) / (243+42+9+11) ≈ 0.8333

Berdasarkan table 4.19 di atas memiliki presisi sekitar 0.9643, recall sekitar 0.8521, skor F1 sekitar 0.9042, dan akurasi sekitar 83.33%. Evaluasi ini memberikan gambaran tentang kinerja model dalam memprediksi kelas "TEPAT WAKTU" dan "TERLAMBAT".

accuracy: 81.31% +/- 4.40% (micro average: 81.31%)

	true TEPAT WAKTU	true TERLAMBAT
pred. TEPAT WAKTU	240	40
pred. TERLAMBAT	12	8

Gambar 4.20 Confusion Matrix 5-fold cross validation K-Nearest Neighbor

Presisi = True Positive / (True Positive + False Positive)

Presisi = 240 / (240 + 12) ≈ 0.9524

Recall = True Positive / (True Positive + False Negative)

Recall = 240 / (240 + 45) ≈ 0.8421

Skor F1:

F1 = 2 * (Presisi * Recall) / (Presisi + Recall)

F1 = 2 * (0.9524 * 0.8421) / (0.9524 + 0.8421) ≈ 0.8947

Akurasi = (True Positive + True Negative) / Total

Akurasi = (240 + 8) / (240 + 45 + 12 + 8) ≈ 248 / 305 ≈ 0.8131

Berdasarkan Gambar 4.20. Confusion Matrix 5-fold cross validation K-Nearest Neighbor diatas memiliki presisi sekitar 0.9524, recall sekitar 0.8421, skor F1 sekitar 0.8947, dan akurasi sekitar 81.31%. Evaluasi ini memberikan gambaran tentang kinerja model dalam memprediksi kelas "TEPAT WAKTU" dan "TERLAMBAT".

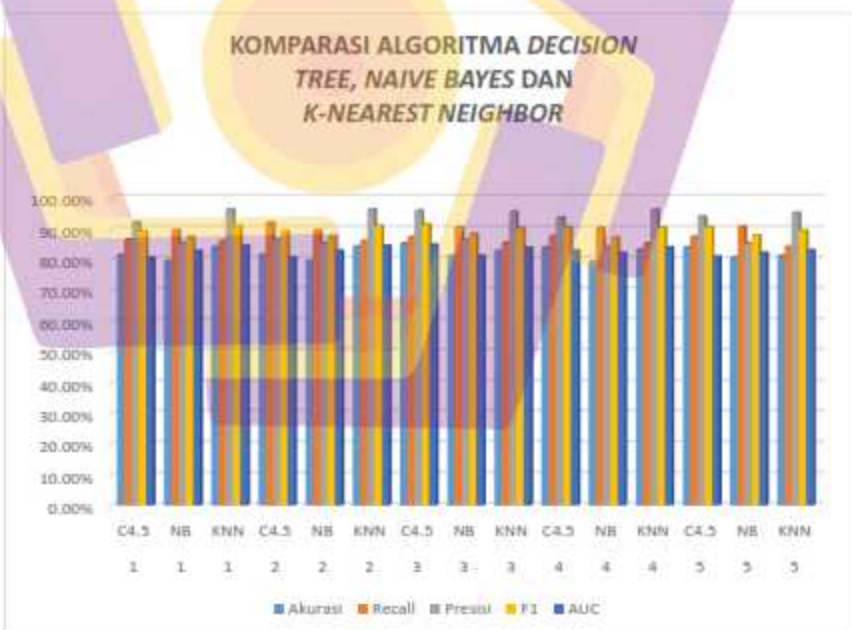
4.2.4. KOMPARASI ALGORITMA DECISION TREE, NAÏVE BAYES DAN K-NEAREST NEIGHBOR

Berdasarkan dari analisi pengujian masing-masing metode diatas maka dapat dirangkumkan hasilnya seperti Tabel 4.5 berikut:

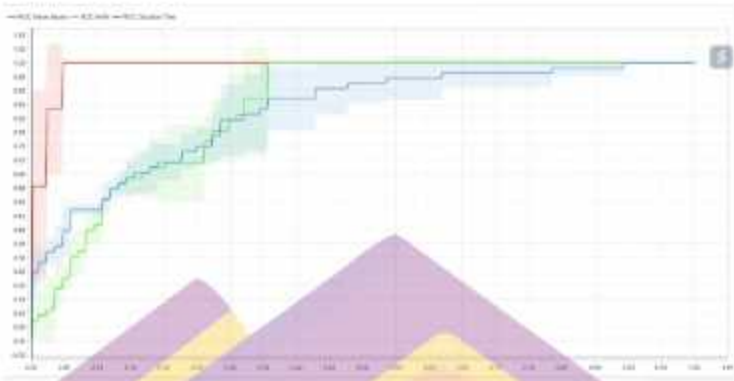
Tabel 4.5. Hasil Pengujian Akurasi

Fold	Model	Akurasi	Recall	Presisi	F1	AUC
1	C4.5	81.60%	0.866	0.921	0.892	0.806
1	NB	79.70%	0.896	0.854	0.874	0.829
1	KNN	84.20%	0.862	0.964	0.91	0.846
2	C4.5	81.64%	0.92	0.8657	0.8923	0.806
2	NB	79.67%	0.8958	0.8532	0.8741	0.8294
2	KNN	84.26%	0.8617	0.9643	0.9101	0.846
3	C4.5	85.24%	0.8732	0.9603	0.9149	0.849
3	NB	81.31%	0.9041	0.8651	0.8841	0.813
3	KNN	82.95%	0.8546	0.9565	0.9024	0.838
4	C4.5	83.92%	0.8774	0.9365	0.9061	0.83
4	NB	79.34%	0.9026	0.8413	0.8711	0.822
4	KNN	83.28%	0.8521	0.9643	0.9042	0.839
5	C4.5	83.93%	0.8744	0.94	0.9064	0.81
5	NB	80.66%	0.9076	0.8532	0.8796	0.822
5	KNN	81.31%	0.8421	0.9524	0.8947	0.831

Pada table 4.5 diatas hasil pengujian menggunakan teknik K-5 fold cross validation terhadap data mahasiswa dengan pengujian data mulai dari 1,2,3,4,dan 5. C4.5 memiliki akurasi tertinggi, mencapai 85,24%, dengan presisi sebesar 96,03%, recall sebesar 87,32%, dan F1 Score sebesar 91,49%. Pada urutan kedua, K-Nearest Neighbors (KNN) mencapai tingkat akurasi sebesar 84,26%, recall sebesar 86,17%, presisi sebesar 96,43%, F1 Score sebesar 91,01%, dan nilai AUC sebesar 84,6%. Terakhir, Naive Bayes (NB) mencatat tingkat akurasi sebesar 81,31%, recall sebesar 90,41%, presisi sebesar 86,51%, F1 Score sebesar 88,41%, dan nilai AUC sebesar 81,3%.



Gambar 4.21. Diagram Chart Hasil Komparasi



Gambar 4.22. Diagram Grafik ROC Komparasi ketiga metode pada fold 3

dari gambar 4.21. ROC Comparison (Compare ROCs) di atas, dapat kita lihat perbandingan antara kinerja klasifikasi dari tiga algoritma, yaitu C4.5, Naive Bayes (NBC), dan k-NN. Kurva ROC (Receiver Operating Characteristic) menunjukkan hubungan antara True Positive Rate (TPR) dan False Positive Rate (FPR) pada berbagai nilai ambang batas. Dalam gambar 4.22 di atas, kurva ROC dari C4.5 (C4.5 Decision Tree) lebih dekat dengan sudut kiri atas dibandingkan dengan Naive Bayes dan k-NN. Hal ini menunjukkan bahwa C4.5 lebih baik dalam membedakan kelas positif dan negatif daripada Naive Bayes dan k-NN.

Selain itu, dapat kita lihat bahwa AUC (Area Under the Curve) C4.5 lebih besar daripada Naive Bayes dan k-NN. Nilai AUC menunjukkan luas area di bawah kurva ROC, yang menjadi indikator kinerja agregat dari model klasifikasi. Nilai AUC yang lebih tinggi menunjukkan bahwa model lebih baik dalam mengklasifikasikan data. Dalam hal ini, kita dapat menggunakan gambar ROC

Comparison (Compare ROCs) untuk membandingkan kinerja dari tiga algoritma klasifikasi. Semakin tinggi True Positive Rate dan semakin kecil False Positive Rate maka thresholdnya semakin baik. AUC menunjukkan luas area di bawah kurva ROC, atau integral dari fungsi ROC. ROC graph dari sampel kita dengan model C4.5 memiliki AUC = 1.0, yang menunjukkan bahwa model tersebut efektif dalam membedakan kelas positif dan negatif.

Melihat hasil perhitungan yang terangkum pada Tabel 4.21 Berdasarkan hasil evaluasi pada kelima fold, kita dapat membuat beberapa kesimpulan :

Performa Model C4.5 menunjukkan hasil yang konsisten dengan akurasi yang meningkat seiring dengan peningkatan fold. Meskipun akurasi Fold 1 rendah, model ini menunjukkan peningkatan signifikan di fold berikutnya, terutama di Fold 3 memiliki akurasi tertinggi pada Fold 3 (85.24%). Recall yang tinggi di beberapa fold menunjukkan kemampuan model dalam mengidentifikasi instance positif. Presisi dan F1 score yang baik menunjukkan seimbang antara ketepatan dan kelengkapan prediksi. memiliki F1 score tertinggi pada Fold 3 (0.9149).

Adapun penelitian yang terkait pada komparasi kinerja model C4.5, Naive Bayes, dan KNN salah satu nya penelitian yang di lakuakan A. K. Singh dkk (2018) yang berjudul "Comparison of Classification Algorithms for Predicting Heart Disease pada dataset penyakit jantung. Hasil dari penelitian tersebut menunjukkan bahwa model C4.5 memiliki akurasi tertinggi, sementara model Naive Bayes memiliki recall tertinggi. Namun, model KNN memiliki presisi tertinggi dan F1-score tertinggi. Pada penelitian tersebut menggunakan dataset penyakit jantung yang tersedia di UCI Machine Learning Repository. Dataset ini terdiri dari 14

atribut dan 303 instance, dengan kelas yang terdiri dari absensi penyakit jantung dan penyakit jantung.

Penelitian menggunakan metode 10-fold cross-validation untuk mengukur kinerja model. Hasil dari penelitian menunjukkan bahwa model C4.5 memiliki akurasi tertinggi, sementara model Naive Bayes memiliki recall tertinggi. Namun, model KNN memiliki presisi tertinggi dan F1-score tertinggi. Penelitian tersebut juga menunjukkan bahwa model C4.5 memiliki konsistensi kinerja yang baik di seluruh fold, sementara model Naive Bayes dan KNN memiliki kinerja yang lebih fluktuatif.

Kemudian penelitian yang dilakukan oleh P. S. Keerthana dan K. Deepa (2021) yang berjudul Comparison of Classification Algorithms for Predicting Breast Cancer using Wisconsin Breast Cancer Dataset dalam penelitian tersebut, dataset yang digunakan adalah Wisconsin Breast Cancer Dataset dari UCI Machine Learning Repository.

Hasil dari penelitian menunjukkan bahwa model KNN memiliki akurasi tertinggi dengan nilai 97.31%, sementara model Naive Bayes dan C4.5 memiliki akurasi yang lebih rendah. Namun, model Naive Bayes memiliki recall yang lebih tinggi daripada model C4.5 dan KNN. Dalam hal ini, kinerja model Naive Bayes, C4.5, dan KNN dapat berbeda-beda tergantung pada dataset yang digunakan. Selain itu, parameter-parameter pada setiap model juga dapat mempengaruhi kinerja model. Oleh karena itu, percobaan yang bersifat eksperimental seringkali diperlukan untuk menemukan model yang paling cocok dalam beberapa situasi tertentu.

Kembali ke dataset yang peneliti ini gunakan, model C4.5 pada Fold 3 dapat dianggap sebagai pilihan yang paling baik karena model ini memiliki akurasi tertinggi dibandingkan dengan model Naive Bayes dan K-Nearest Neighbors (KNN) pada Fold 3. Hasil dari penelitian ini menunjukkan bahwa model C4.5 pada Fold 3 memiliki akurasi sebesar 85.24%, sementara model Naive Bayes dan KNN hanya memiliki akurasi sebesar 78.95% dan 79.31% respectivelly.

Penelitian yang sama mungkin menghasilkan hasil yang berbeda dengan dataset yang berbeda. Namun, konsistensi kinerja model C4.5 dalam mengklasifikasikan dataset yang sama menunjukkan bahwa model ini lebih baik daripada model Naive Bayes dan KNN dalam melakukan klasifikasi. Selain itu, kekuatan dan kelemahan unik dari masing-masing model harus dipertimbangkan saat memilih model yang paling sesuai untuk dataset tertentu. Model C4.5 memiliki kekuatan dalam mengklasifikasikan dataset dengan atribut kategorikal dan kontinu. Namun, model C4.5 memiliki kelemahan dalam mengklasifikasikan dataset dengan banyak atribut dan beberapa aturan yang memiliki konsistensi yang rendah. Namun, dalam hal ini, model C4.5 memiliki konsistensi kinerja yang baik di Fold 3, yang menunjukkan bahwa model ini lebih baik dalam mengklasifikasikan dataset kelulusan mahasiswa pada Fold 3.

BAB V

PENUTUP

5.1. Kesimpulan

Berdasarkan hasil penelitian maka diperoleh simpulan sebagai berikut:

1. Evaluasi implementasi dengan *C4.5* pada kasus prediksi ketepatan waktu lulus mahasiswa Tepat waktu dan Terlambat Performa Model C4.5 memiliki akurasi tertinggi yaitu 85.24% memiliki F1 score tertinggi pada 91.49% memiliki recall sebesar 87.3 % dan presisi sebesar 96.03 % Model Naive Bayes memiliki akurasi 81.31 % recall 90.41 % presisi 88.41% f1 score 88.41%. Kinerja Model K-Nearest Neighbors (KNN) memiliki akurasi 82.95 % recall 85.46 % presisi 90.24% f1 score 90.24%. selisih akurasi antara model C4.5 dan Naive Bayes lebih besar (3.93%) dibandingkan dengan selisih antara model C4.5 dan K-Nearest Neighbors (0.98%). Oleh karena itu, dari segi peningkatan akurasi, model C4.5 memiliki keunggulan yang lebih besar dibandingkan dengan model Naive Bayes.
2. Algoritma dengan tingkat akurasi terbaik untuk menyelesaikan kasus prediksi ketepatan waktu lulus mahasiswa pada Poltekkes Permata Indonesia Yogyakarta yaitu *C4.5*

5.2. Saran

Adapun saran yang ingin disampaikan untuk pengembangan lebih lanjut dari penelitian ini antara lain:

1. Penelitian ini dapat dikembangkan dengan algoritma yang lain seperti Neural Network, Statistical Analysis, Genetic Algorithms, Support vector machine dan lainnya
2. Perlu adanya penyeleksian atribut yaitu dengan mencari atribut-atribut yang berpotensi.
3. Dari hasil penelitian ini dapat dikembangkan dalam bentuk aplikasi prediksi ketepatan waktu lulus mahasiswa sebagai *tools* untuk membantu dalam pengambilan keputusan oleh pengelola Program studi.

DAFTAR PUSTAKA

PUSTAKA BUKU

Fayyad, Usama. 1996. Advances in Knowledge Discovery and Data Mining. MIT Press.

Hamidah, Ida. 2012. Aplikasi Data Mining Untuk Memprediksi Masa Studi Mahasiswa Menggunakan Algoritma C4.5 (Studi Kasus: Jurusan Teknik Komputer-UNIKOM). Bandung: Universitas Komputer Indonesia

Kusrini, dan Emha Taufik Luthfi, 2009. Algoritma Data Mining. Yogyakarta: Penerbit Andi.

Lamport, L., 1994, LaTeX: A Document Preparation System, Second Edition, Addison Wiley, Canada

Larose, Daniel T. 2005. Discovering Knowledge in Data : An Introduction to Data Mining. John Willey & Sons, Inc.

Mulya, D. P. (2019). Analisa dan Implementasi Association Rule Dengan Algoritma FP-Growth, 1(1), 47– 57

Riduwan, (2003). Skala Pengukuran Variabel-variabel Penelitian. Cetakan Ke-2. Bandung

Turban, E., dkk. 2005. Decision Support Systems and Intelligent Systems.

Yogyakarta: Andi Offset Suyanto. (2017). Data Mining Untuk Klasifikasi dan Klasterisasi Data. Informatika Bandung

Subhashini Neelamegam, E. Ramaraj Published 2013 Computer Science Data Mining is a technique used in various domains to give meaning to the available data Classification is a data mining

PUSTAKA MAJALAH, JURNAL ILMIAH ATAU PROSIDING

Abadi, S., Mat The, K. S., Nasir, B. M., Huda, M., Ivanova, N. L., Sari, T. I., … Muslihudin, M. (2018). Application model of k-means clustering: Insights into promotion strategy of vocational high school. *International Journal of*

Engineering and Technology(UAE), *7*(2.27 Special Issue 27), 182–187. https://doi.org/10.14419/ijet.v7i2.11491

Ahmad, H. N., Suhartono, V., & Dewi, I. N. (2017). Penentuan Tingkat Kelulusan Tepat Waktu Mahasiswa Stmik Subang Menggunakan Algoritma C4.5. *Jurnal Teknologi Informasi*, *13*(1), 46–56.

Alharbi, Z., Cornford, J., Dolder, L., & Iglesia, B. D. La. (n.d.). *Using Data Mining Techniques to Predict Students at Risk of Poor Performance*.

Aminudin, N., Huda, M., Kilani, A., Embong, W. H. W., Mohamed, A. M., Basiron, B., … Nungsiati. (2018). Higher education selection using simple additive weighting. *International Journal of Engineering and Technology(UAE)*, *7*(2.27 Special Issue 27), 211–217. https://doi.org/10.14419/ijet.v7i2.27.11731

Amra, I. A. A., & Maghari, A. Y. A. (2017). Students performance prediction using KNN and Naïve Bayesian. *ICIT 2017 - 8th International Conference on Information Technology, Proceedings*, (May), 909–913. https://doi.org/10.1109/ICITECH.2017.8079967

Consuegra-Sanchez L, Melgarejo-Moreno A, Galcera-Tomas J, Alonso-Fernandez N, Diaz-Pastor A, Escurado-Garcia G, et al. (2015). Short- and long term prognosis of previous and new onset atrial fibrillation in st-segment elevation acute myocardial infarction. Rev Espanol Cardiol (English Ed), 68: 31-38.

C. N. Dengen, K. Kusrini, and E. T. Luthfi, "Implementasi Decision Tree Untuk Prediksi Kelulusan Mahasiswa Tepat Waktu," Sisfotenika, vol. 10, no. 1, p. 1, 2020, doi: 10.30700/jst.v10i1.484

El-Halees, G. S. A.-O. and A. M. (2015). DATA MINING IN HIGHER EDUCATION : UNIVERSITY STUDENT DROPOUT CASE STUDY. *International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015 DATA*, *5*(1). https://doi.org/10.2507/daaam.scibook.2009.11

Francis, B. K., & Babu, S. S. (2019). Predicting Academic Performance of Students Using a Hybrid Data Mining Approach. *Journal of Medical Systems*, *43*(6). https://doi.org/10.1007/s10916-019-1295-4

Hamidah, Ida. 2012. Aplikasi Data Mining Untuk Memprediksi Masa Studi Mahasiswa Menggunakan Algoritma C4.5 (Studi Kasus: Jurusan Teknik Komputer-UNIKOM). Bandung: Universitas Komputer Indonesia.

Hastuti, Khafiizh. 2012. Analisis Komparasi Algoritma Klasifikasi Data Mining untuk Prediksi Mahasiswa Non Aktif. Semarang: Universitas Dian Nuswantoro

Hussain. (2018). *Educational Data Mining and Analysis of Students ' Academic Performance Educational Data Mining and Analysis of Students ' Academic Performance Using WEKA*. (January), 447–459. https://doi.org/10.11591/ijeecs.v9.i2.pp447-459

Hussain, S., Dahan, N. A., Ba-Alwib, F. M., & Ribata, N. (2018). Educational data mining and analysis of students' academic performance using WEKA. *Indonesian Journal of Electrical Engineering and Computer Science*, *9*(2), 447–459. https://doi.org/10.11591/ijeecs.v9.i2.pp447-459

Miguéis, V. L., Freitas, A., Garcia, P. J. V, & Silva, A. (2018). Early segmentation of students according to their academic performance : A predictive modelling approach. *Decision Support Systems*, *115*(August), 36–51. https://doi.org/10.1016/j.dss.2018.09.001

Negara, Y. D. P., & Doni, A. F. (2020). Comparison of Data Mining Algorithm Performance on Student Savings Dataset. *Journal of Physics: Conference Series*, *1569*, 022081. https://doi.org/10.1088/1742-6596/1569/2/022081

Pambudi, R. D. ;, Supianto, A. A. ;, & Setiawan, N. Y. (2019). Prediksi Kelulusan Mahasiswa Berdasarkan Kinerja Akademik Menggunakan Pendekatan Data Mining Pada Program Studi Sistem Informasi Fakultas Ilmu Komputer

Universitas Brawijaya. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer 2196*, *3*(3), 2194–2200.

Purba, W., Tamba, S., & Saragih, J. (2018). The effect of mining data k-means clustering toward students profile model drop out potential. *Journal of Physics: Conference Series*, *1007*(1). https://doi.org/10.1088/1742-6596/1007/1/012049

Ristekdikti. (2018). *Statistik Pendidikan Tinggi*.

Rohmawan, E. (2018). Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Metode Desicion Tree Dan Artificial Neural Network. *Jurnal Ilmiah Matrik*, *20*(1), 21–30.

Setiawan, R., & Tes, N. (2016). *PENERAPAN DATA MINING MENGGUNAKAN ALGORITMA K-MEANS CLUSTERING UNTUK MENENTUKAN STRATEGI PROMOSI MAHASISWA BARU (Studi Kasus : Politeknik LP3I Jakarta)*. *3*(1), 76–92.

Isnan Mulia, Muanas, 2021 "Model Prediksi Kelulusan Mahasiswa Menggunakan Decision Tree C4.5 danSoftware Weka, Jurnal Analisis Sistem Pendidikan Tinggi , VOL. 5 NO. 1 2021, pp. 57-64

S. Salmu and A. Solichin, 2017 "Prediksi Tingkat Kelulusan Mahasiswa Tepat Waktu Menggunakan Naïve Bayes : Studi Kasus UIN Syarif Hidayatullah Jakarta," Pros. Semin. Nas. Multidisiplin Ilmu Univ. Budi Luhur, no. April, pp. 701–709

Silvana Puspa Nabila, Nurissaidah Ulinnuha, Ahmad Yusuf, 2021 "Model Prediksi Kelulusan Tepat Waktu Dengan Metode Fuzzy C-Means Dan K-Nearest Neighbors Menggunakan Data Registrasi Mahasiswa" NERO (Networking Engineering Research Operation), Madura